On the Relation Between Damage and Deception

Joel Sobel

August 29, 2023

Three Tasks

1. "Pedagogic": Provide coherent definition(s) of deception.
2. "Applied": Relate the definition to examples faced by regulators.
3. "Technical": Propositions that relate binary relationships on beliefs to preferences of decision makers.

Framework

- One Decision Maker.
- Beliefs $\mu \in \Delta(\Theta)$.
- Decision maker takes action y.
- Utility function: $U^R(\theta, y)$.
- For convenience: Θ finite.

Standard Interpretation

Two players: Sender and Receiver.

Sender observes $\theta \in \Theta$.

Sender sends message $m \in M$. (M can depend on θ.)

R observes m.

Receiver takes action $y \in Y$.

Preferences $U^i(\theta, y, m)$; $U^R(\cdot)$ independent of m.

Prior $P(\theta)$ (positive on Θ).

Explanation

Standard interpretation

1. . . . introduces game theory.
2. . . . explains where beliefs come from.
3. . . . relevant for applications.

Deception: Informal

- Deception is "inducing bad beliefs."

  I need some notion of beliefs.

  I need some notion of "bad."

- How to do this?

  Let $D(\mu)$ is the set of beliefs that are less accurate/more deceptive than μ.

- So beliefs are bad compared to other possible beliefs.

Easy Case: Two states. Beliefs totally ordered (by probability placed on true state).

Conceptual Approach: Deception

- Deception is partial order on beliefs. Loosely "more deceptive" means "further from truth."
  (μ more deceptive than $\mu' \ldots$)
- There will be lots of definitions because there are lots of distances from truth.
- To evaluate the relevance of a particular definition, I relate it to when "bad" beliefs lead to "bad" utility.

Conceptual Approach: Damage

- Damage is a(nother) partial order on beliefs. Loosely "more damaging" means "leads to lower utility."
- There will be lots of definitions because there are lots different kinds of agent/utility function.
- Goal: "Damage-Deception Result" associating a definition of deception with a class of preferences.

Strong (S) Deception

Definition (Strong deception)

The belief μ' is more **strongly deceptive** than μ given θ^* if $\mu' \neq \mu$ and there exists $p \in [0, 1)$ such that

$$\mu(\cdot) = p\mu'(\cdot) + (1 - p)I(\cdot \mid \theta^*). \quad (1)$$

μ on segment connecting point mass on θ^* to μ'.

[$I(\cdot \mid \theta^*)$ is point mass on θ^*.]

S-Deception Illustrated

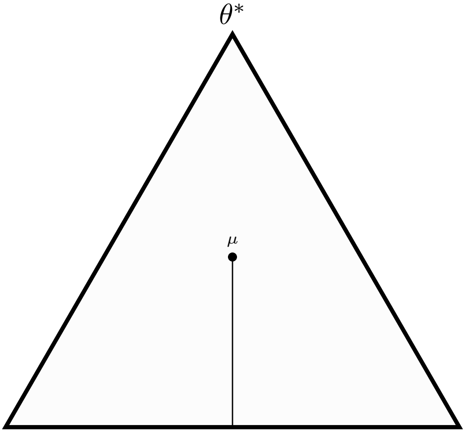

Figure: Beliefs on the line segment are more strongly deceptive than μ.

S-Deception Illustrated Again

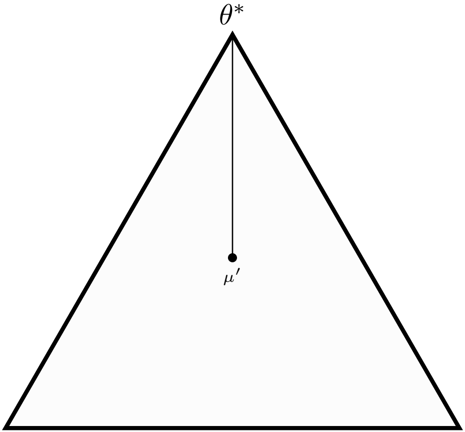

Figure: The belief μ' is more strongly deceptive than μ if only if μ is on line segment.

Damage: Preliminaries

Deception: Property of Beliefs

Damage: Consequences. Property of Preferences.

Assume R best replies to beliefs.

$BR(\mu)$ is R's best response correspondence:

$$BR(\mu) = \arg\max_{y \in Y} \sum_{\theta \in \Theta} U^R(\theta, y)\mu(\theta).$$

Damage: Definition

Let $\bar{u}(\theta, \mu) = U^R(\theta, BR(\mu))$.

Definition (Damaging Behavior)

The belief μ' is more **damaging** than μ given θ^* if

$$\bar{u}(\theta^*, \mu') < \bar{u}(\theta^*, \mu).$$

Technicality:

- If $BR(\mu)$ is not single valued, then $\bar{u}(\theta^*, \mu')$ may not be single valued.
- So $\bar{u}(\theta^*, \mu') < \bar{u}(\theta^*, \mu)$ needs a definition.
- I need to make some assumption (otherwise ugly statements of propositions).
- In what follows, can rank sets using strong set order, maximum, minimum, ...

Damaging Relative to a Set

Let $\mathcal{U}$ be a set of payoff functions for R (real valued functions of (θ, y)).

Definition (Damaging Relative to a Set)

The belief μ' is more **damaging relative to** $\mathcal{U}$ than μ given θ^* if

1. $\bar{u}(\theta^*, \mu') \leq \bar{u}(\theta^*, \mu)$ for all $u \in \mathcal{U}$
2. $\bar{u}(\theta^*, \mu') < \bar{u}(\theta^*, \mu)$ for some $u \in \mathcal{U}$.

The same technicality about "$<$" applies.

Where am I going?

1. $\mathcal{U}$ generates partial order on beliefs.
2. I want to compare these partial orders to definitions of deception.

Aside: FTC

The Federal Trade Commission identifies three necessary conditions for deception.

1. Deception requires doing something that misleads the consumer.
2. FTC evaluates the impact from the perspective of a consumer who acts reasonably.
3. For a practice to be deceptive it must have a material impact on a consumer.

FTC versus Me

- My notion of deception concentrates on “misleading” information.
- My formal results connect “material impact” (damage) to misleading information.

Generic Proposition Statement

Proposition

The belief μ' is more X deceptive than μ given θ^* if and only if it is more damaging than μ relative to the $F(X)$ family of preferences given θ.

The talk provides specific $(X, F(X))$ pairs.

Teaser:
When $X =$ strong deception, $F(X) =$ all preferences.

Haven't I seen this before?

1. μ FOSD μ' if and only if all decision makers with increasing utility functions prefer μ.
2. μ SOSD μ' if and only if all decision makers with concave utility functions prefer μ.

Contrast

My approach is interim, not ex ante.

My approach involves a decision.

Persuasion Framework

- R's decision is whether to accept or reject a proposal.
- $Y = \{0, 1\}$. $y = 0$ reject; $y = 1$ accept.
- $A^* = \{\theta : U^R(\theta, 1) \geq U^R(\theta, 0)\}$. Acceptable states for R.
- Persuasion preferences (with respect to A^*):

Definition (Persuasion Preferences)

R has **persuasion preferences** if $U^R(\theta, 0) = 0$ and

$$U^R(\theta, 1) = \begin{cases} W & \text{if } \theta \in A^* \\ -L & \text{if } \theta \notin A^* \end{cases}$$

for $W, L > 0$.

Persuasion Result

Definition (Binary-Action Deception)

The belief μ' is more **binary-action (BA)-deceptive** than μ given θ^* if

$$\text{if } \theta^* \in A^*, \text{ then } \mu'(A^*) < \mu(A^*)$$
$$\text{and if } \theta^* \notin A^*, \text{ then } \mu'(A^*) > \mu(A^*).$$

Proposition

The belief μ' is more binary-action-deceptive than μ given θ^* if and only if it is more damaging than μ given θ^* relative to the family of persuasion preferences.

When $X =$ binary-action deception, $F(X) =$ persuasion preferences.

S-Deception Result

Proposition

The belief μ' is more strongly deceptive given θ^* than μ if and only if it is more damaging than μ given θ^* relative to the set of all preferences.

$X =$ strong deception
$F(X) =$ all preferences

Proportional (P) Deception

Definition (Proportional Deception)

The belief μ' is more **proportional (P)-deceptive** than μ given θ^* if $\mu(\theta^* \mid n) > 0$, and there exists a number $p \in [0, 1)$, and a distribution ρ satisfying $\rho(\theta^*) = 0$ such that

$$\mu'(\cdot) = p\mu(\cdot) + (1 - p)\rho. \tag{2}$$

Equivalent to

$$\frac{\mu(\theta^*)}{\mu(\theta)} \geq \frac{\mu'(\theta^*)}{\mu'(\theta)}$$

(θ^* relatively more likely under n.)

P-Deception Illustrated

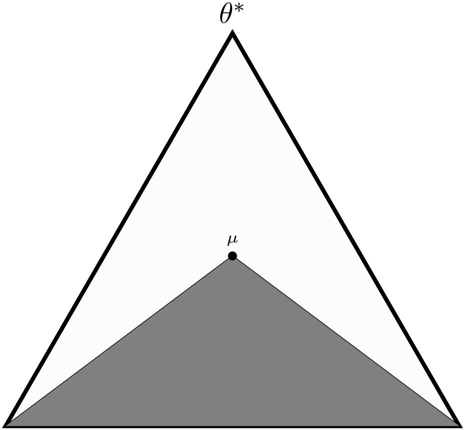

Figure: Beliefs in shaded region are more proportionally deceptive than μ given θ^*.

Again . . .

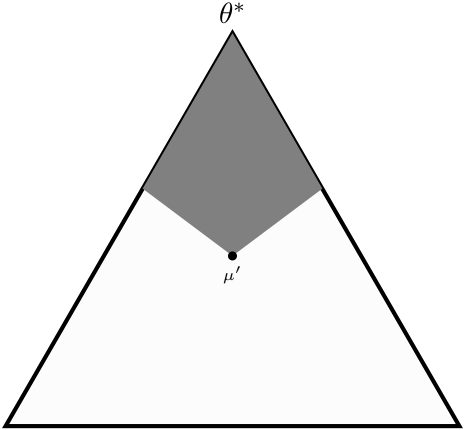

Figure: The belief μ' is more proportionally deceptive than μ if and only if μ is in shaded region.

P-Deception Result

Definition (State Specific)

The Receiver's preferences are **state specific** if there is a bijection $\phi : \Theta \to Y$ and positive numbers $\alpha(\theta)$ for $\theta \in \Theta$ such that

$$U^R(\theta, y) = \begin{cases} \alpha(\theta) & \text{if } y = \phi(\theta) \\ 0 & \text{if } y \neq \phi(\theta) \end{cases}$$

Corresponds to situation in which there is a correct action for each state.

Proposition

The belief μ' is more proportionally deceptive than μ given θ^* if and only if it is more damaging than μ given θ^* relative to the family of state-specific preferences.

Variations

1. Constrained
2. Kullback–Leibler

Ordered states: Monotone

Nesting Illustrated

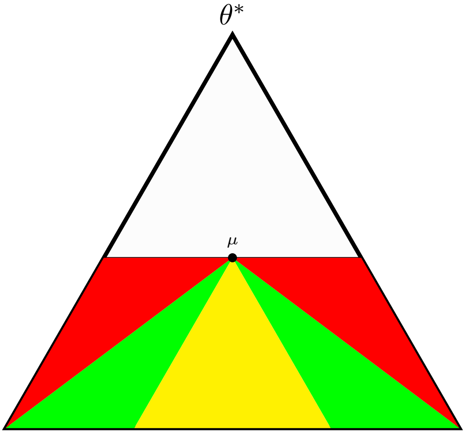

Figure: Relative to μ and given θ, beliefs in red region are KL-deceptive but not P-deceptive. Beliefs in green region are P-deceptive but not C-deceptive. Beliefs in green region are C-deceptive but not S-deceptive.

Strategic Considerations

1. Imagine S sends messages, messages determine beliefs.
2. Look for equilibria of persuasion game.
3. Define partial order on messages based on beliefs/actions induced.

   m' is more deceptive than m if and only if $\mu(\cdot \mid m')$ is more deceptive than $\mu(\cdot \mid m)$.

   m' is deceptive if there exists m such that m' is more deceptive than m.
4. Must specify S preferences.

Comments

Cases

1. R can avoid being deceived. [No deception/damage if R's beliefs and actions don't depend on S's message.]
2. R may prefer (ex ante) an eq in which he is deceived.
3. If S always wants to persuade, talk is cheap, prior is favorable, and some message induces $y = 0$, then damage and deception.
4. If S sometimes does not want to persuade, then some message will induce $y = 0$.

Practical Considerations

Definition of deception depends on context:

1. Persuasion Preferences: R decides whether to buy.
2. State Specific Preferences: R decides which product to buy.
3. Monotone Preference: R decides how much to buy.
4. Strong Deception: Conservative.

Comments on Welfare

1. Large penalties for deception benefit R because R can "force" fully revealing outcome.
2. This property depends on equilibrium selection.
3. Large penalties for deception may benefit R and S.
4. ... but not if one selects babbling when an ex ante superior equilibrium exists.
5. Restrictions on deception may harm both if it is costly to disclose (by driving firms out of business).

Theoretical Properties

1. For which binary relations is there a damage-deception result?
2. Minimal relation?
3. Maximal relation?

Back to the Three Tasks

1. "Pedagogic": Mission accomplished (well, too many definitions)
2. "Applied": Only suggested in this talk.

   Messages:

   - Proper definition of deception depends on the context (what is known about R's preferences).
   - Deception is possible in equilibrium.
   - An R who can be deceived may also be an R who may benefit from communication.
3. "Technical": Dam-Dec results. Deception relation more complete corresponds to smaller families of associated preferences.

   Harming R is a consequence of deception rather than part of the definition.

Tribute

gpt

Supplementary Material

Follows

General Properties

- There exist (non-trivial) deception correspondences.
- The deception relations are transitive.

More Properties

Deception Correspondences

- Deception relations are not symmetric:

Proposition

If D is a deception correspondence with respect to $\mathcal{U}$ given θ^*, then $\mu' \in D(\mu'')$ implies $\mu'' \notin D(\mu')$.

- Deception correspondences are not convex valued in general, but

Proposition

If D is a deception correspondence with respect to $\mathcal{U}$ given θ^*, $\mu \in D(\mu^*)$, and $\mu = p\mu' + (1-p)I(\cdot \mid \theta^*)$ for $p \in (0,1]$ then $\mu' \in D$.

Constrained (C) Deception

Constrained

Definition (Constrained deception)

The belief μ' is more **constrained (C)-deceptive** than μ given θ^* if $\mu(\theta^*) > 0$ and either

$$\mu'(\theta^*) = 0$$

or there exists a number $p \in [0, 1)$, a distribution ρ satisfying $\rho(\theta^*) = 0$ such that

$$\mu'(\cdot) = p\mu(\cdot) + (1 - p)\rho$$

and

$$\rho(\theta) \geq \mu(\theta) \quad \text{for all } \theta \neq \theta^*.$$

μ' lowers probability of true state and raises it on others.

C-Deception Illustrated

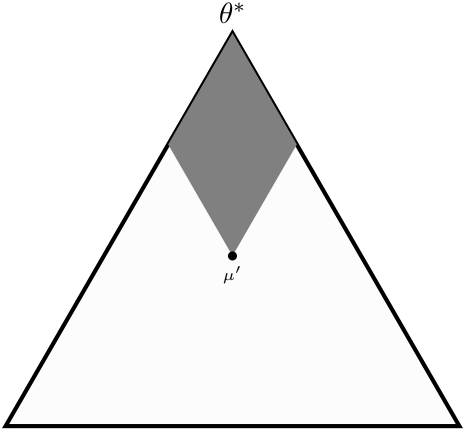

Figure: The belief μ' is more constrained deceptive than μ if and only if μ is in the shaded region.

C-Deception Again

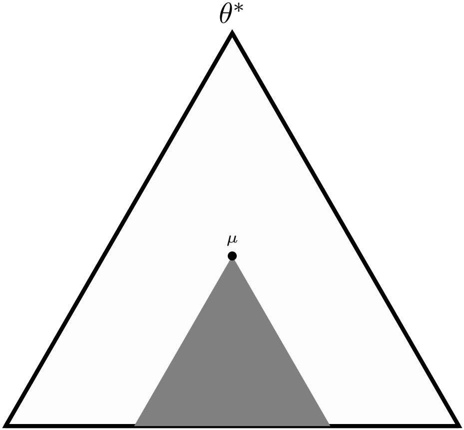

Figure: Beliefs in shaded region are more constrained deceptive than μ given θ^*.

C-Deception Result

Definition (Linear Family with outsider option)

If R has preferences in the family of linear preferences with an outside option u_0, then R selects an action $y \in \Delta(\Theta \cup [0, 1])$ to maximize

$$\sum_{\theta} \mu(\theta) \frac{y(\theta)}{\beta(\theta)} + y_0 u_0.$$

Proposition

The belief μ' is more constrained deceptive than μ given θ^* if and only if it is more damaging than μ relative to the family of linear preferences with an outside option given θ^*.

Nesting

Kullback-Leibler (KL) Deception

Kullback–Leibler divergence between distributions μ and μ':

$$D_{KL}(\mu' \mid\mid \mu) = \sum \mu'(\theta) \log \frac{\mu'(\theta)}{\mu(\theta)}.$$

[Assume that $\mu(\theta) = 0$ implies $\mu'(\theta) = 0$ and follow the convention that $x \log x = 0$ when $x = 0$.]

Definition (KL-Deception)

The belief μ' is more **Kullback-Leibler (*KL*)-deceptive** than μ given θ^* if $\mu(\theta^* > 0$ and

$$D_{KL}(I(\cdot \mid \theta^*) \mid\mid \mu'(\cdot) < D_{KL}(I(\cdot \mid \theta^*) \mid\mid \mu'(\cdot).$$

Simpler

A message is KL-deceptive if and only if there is another message that induces beliefs placing higher probability on the true state.

KL-Deception Illustrated

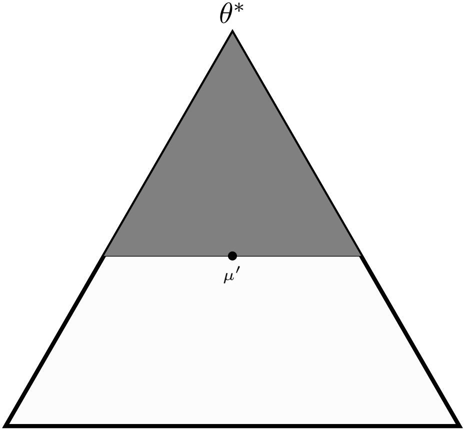

Figure: The belief μ' is more KL-deceptive than μ given θ^* if and only if μ is in shaded region.

Again

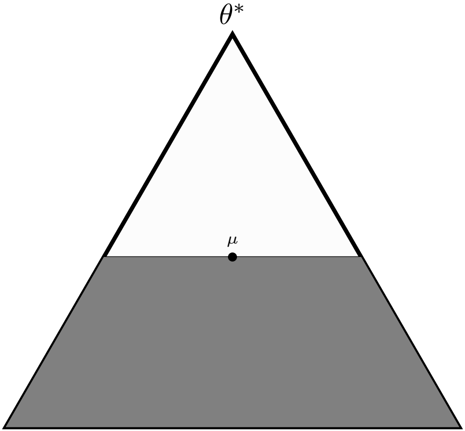

Figure: Beliefs in shaded region are more KL-deceptive than μ given θ^*.

KL-Deception Result

Definition (Exponential Family)

The family of preferences is logarithmic if $g(y) = \log y$.

Proposition

The belief μ' is more Kullback-Leibler deceptive than μ given θ^* if and only if it is more damaging than μ relative to the family of logarithmic preferences given θ^*.

Nesting

Preliminary

1. Y and Θ linearly ordered.
2. Elements of Θ denoted by $\theta_1, \ldots, \theta_N$ where $\theta_i < \theta_j$ if and only if $i < j$.

Definition (Increasing Differences)

Assume that Θ and Y are completely ordered. The function $u : \Theta \times Y \to \mathbb{R}$ satisfies **increasing differences** if $u(\theta_j, y) - u(\theta_i, y)$ is increasing in y whenever $j > i$.

Given a distribution μ, denote by $C(j; \mu)$ the cumulative probability determined by μ. That is, $C(j; \mu) = \sum_{i \leq j} \mu(\theta_i)$.

Definition (First-Order Stochastic Dominance)

μ **dominates** the probability distribution μ' if $\mu \neq \mu'$ and $C(j; \mu') - C(j; \mu) \geq 0$ for all j.

Monotone Deception Defined

Definition (Monotone Deception)

The belief μ' is more **monotonically (M)-deceptive** than μ given θ^* if one of the following conditions hold:

1. μ' is more strongly deceptive than μ given θ^*;
2. μ dominates μ' and $\mu(\theta) = \mu'(\theta) = 0$ for $\theta > \theta^*$;
3. μ' dominates μ and $\mu(\theta) = \mu'(\theta) = 0$ for $\theta < \theta^*$.

Comments

1. The first condition holds and $\mu(\theta)) = \mu'(\theta) = 0$ for $\theta > \theta^*$, then the second condition holds.
2. Likewise for the third condition.
3. Similarly, if the first condition holds and $\mu(\theta) = \mu(\theta) = 0$ for $\theta < \theta^*$, then the third condition holds.
4. So: μ' is more *M*-deceptive than μ if:
    - 4.1 θ^* is neither the highest nor the lowest state in the support of μ' and μ and μ' is more strongly deceptive;
    - 4.2 θ^* is the lowest state given positive probability by μ and μ' is stochastically greater.
    - 4.3 θ^* is the highest state given positive probability by μ and μ' is stochastically lower.

M deception illustrated

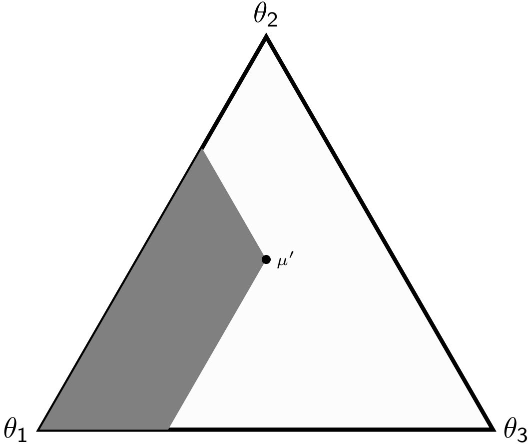

Figure: Belief μ' is more M-deceptive than μ given θ_1 if and only if μ is in shaded region.

M deception illustrated again

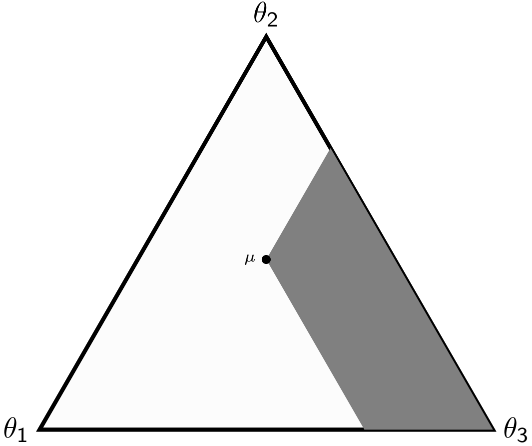

Figure: The shaded area is the set of beliefs that are more M-deceptive than μ given θ_1.

M Deception Result

Proposition

The belief μ' is more monotone deceptive than μ given θ^* if and only if it is more damaging relative to the class of ID preferences.

[*ID* preferences are concave, increasing, and satisfy increasing differences.]

Remarks

1. M deception restrictive (because strong deception is).
2. But: $\theta^* = \theta_1$ case is of particular interest.
   - R would not purchase the item knowing given θ_1.
   - S would like to convince R to buy.
   - Result associates damage with convincing R to buy more than he wants.
   - m is deceptive if it "exaggerates" the true state and there is a more moderate exaggeration available.
3. Different classes of preferences:
   3.1 Tail states don't changes optimal action.
   3.2 Quadratic loss.

Nesting

Zenith Deception

Definition (Zenith (Z) Deception)

The belief μ' is more **zenith deceptive** than μ given θ^* and μ if $\mu'(\theta^*) < \max_\theta \mu'(\theta)$ and $\mu(\theta^) = \max_\theta \mu(\theta)$.

There is an μ makes θ^* most likely (and μ' don't make θ^* most likely).

Z-Deception Illustrated

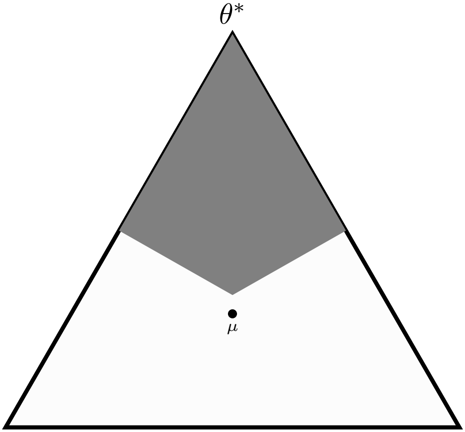

Figure: The belief μ' is more zenith deceptive than μ given θ^* if and only if μ is in the shaded region. ($\mu(\theta^*) = \max_\theta \mu(\theta)$, and $\mu'(\theta^*) < \max_\theta \mu'(\theta)$.)

More . . .

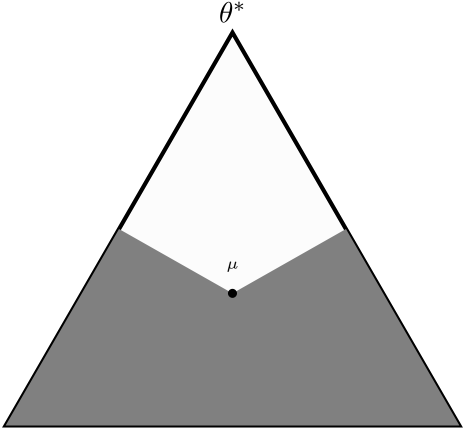

Figure: Beliefs in shaded region are more zenith deceptive than μ given θ^*. ($\mu(\theta^*) = \max_\theta \mu(\theta)$.)

Z-Deception Result

Definition (Uniformly Linear Family)

The family (g, β) of preferences is uniformly linear if $g(\cdot)$ is linear $c > 0$ such that $\beta(\theta) = c$ for all θ.

Proposition

The belief μ' is more zenith deceptive than μ given θ^* if and only if it is more damaging than μ relative to the family of uniform linear preferences given θ^*.

Nesting

Volkswagen

- Volkswagen advertised that they produced diesel cars that were not dangerous to the environment.
- Vehicles were equipped with illegal emission defeat devices during government tests. (Not disclosed.)
- Message: announcements about safety.
- Alternative message: truth, saying nothing.
- Decision: whether to buy.
- Deception: thinking the car was environmentally friendly.
- Damage: buying wrong car in response to the message.

Comments

Machinima

- Machinima paid people who created videos posted on youtube to include Xbox footage in their reviews.
- Machinima asked the youtubers not to reveal payments.
- FTC argued that it was misleading to represent paid endorsers as independent reviewers.
- No evidence that the reviews were false, but FTC claimed withholding information about payments may influence the interpretation of the videos.
- Message: ads
- Alternative: disclosing payments (or not making them)
- Decision: What system to buy
- Deception: Unjustified confidence in quality of system
- Damage: Purchase of wrong system

Comments

POM Wonderful

- The ads claimed that the POM (pomegranate juice) could prevent or reduce the risk of heart disease, prostate cancer, and erectile dysfunction.
- POM provided supporting evidence.
- FTC asserted evidence was inadequate (lacking proper controls; statistically insignificant results).
- Message: ads
- Alternative message: comments on other characteristics; complete description of evidence.
- Decision: whether to buy (or how much to buy).
- Deception: inaccurate impression about health benefits.
- Damage: buying too much.

Comments

Kellogg's Mini-wheats

- Ads claimed that children who ate Frosted Mini-Wheats were 20% more attentive than those who skipped breakfast.
- Kellogg referred to a study to back up the claims.
- FTC argued that claims, while not literally false, were misleading. (Half of subjects showed no increase; only 10% significant gains.)
- Message: ads
- Alternative message: comments on other characteristics; complete description of evidence
- Decision: whether to buy (or how much to buy)
- Deception: inaccurate impression about benefits.
- Damage: buying too much.

Comments

Red Bull

- Red Bull energy drink does **not** give you wings (literally).
- Law suit (and settlement) based on lack of evidence that it gives you (figurative) wings.

Comments

Chat GPT

1. Digital Economics and Platform Markets
2. Behavioral Economics and Nudging
3. Environmental and Resource Economics
4. Health Economics and Healthcare Markets
5. Innovation and Intellectual Property
6. Data Economics and Privacy
7. Development Economics
8. Economics of Information and Learning

Tribute