# Behavioral Environmental Economics

*Theory and Experiments*

Félix Muñoz-García

School of Economic Sciences, Washington State University

Course slides

September 2026

# Roadmap

- The question: how do environmental problems and policies change once people are not the fully rational, self-regarding agents of the textbook?
- Two organizing ideas run through the course. Some departures *enlarge what people value*, and policy works with those wider preferences. Others open an *own-welfare wedge* between what a person chooses and what serves her own welfare, and policy corrects it.
- We take up six behavioral departures in turn. For each one, the model comes first, then the experiments that measure it.
- We then map the mechanisms across environmental domains and turn to the design of corrective policy: prices, standards, and nudges.

# Motivation and the organizing idea

# A puzzle: home energy reports

- A utility mails each household a report comparing its electricity use to that of similar neighbors, with an approving or disapproving symbol
- These reports have gone to more than six million U.S. households
- On average, electricity use falls by about two percent, an effect comparable to a short-run price increase
- Yet no price changed and no budget constraint was relaxed: only the social comparison did

Why is this a puzzle? A standard forward-looking consumer, facing the same budget and prices, has no reason to change. Something outside the textbook model is moving behavior.

Allcott (2011); Allcott and Rogers (2014).

# Puzzles the standard model strains to explain

- Households pay a premium for green electricity and voluntary carbon offsets, though a purely self-regarding calculation says they should not
- People demand far more to give up an environmental amenity than they will pay to obtain the same amenity (the willingness-to-accept versus willingness-to-pay gap)
- Investments in energy efficiency that appear to pay for themselves are still passed over (the "energy-efficiency gap")
- Monetary incentives sometimes *reduce* the very behavior they were meant to reward, by crowding out moral motivation

Each puzzle points to a specific behavioral departure, and we will build a model for each one in turn.

# Two phases of the field

- *Phase 1* (roughly 2000–2010): largely conceptual and cautionary. Should laboratory anomalies, the endowment effect, hyperbolic discounting, cooperation above zero, change environmental valuation and policy at all?
- The reasonable objection: anomalies might not survive real stakes, learning, and markets
- *Phase 2* (from about 2010): large field experiments and structural estimates from market data show several departures are present *at scale*, while also revealing how heterogeneous and context-dependent they are

This course reflects both: the models formalize the phase-1 mechanisms, and the experiments supply the phase-2 magnitudes.

# Earlier surveys, and our angle

- Prior surveys catalogue the field by behavioral assumption or by application: Shogren–Taylor (2008); Croson–Treich (2014); Gsottbauer–van den Bergh (2011); Carlsson–Johansson-Stenman (2012); Kesternich et al. (2017); recently Rodemeier–Semken (2026)
- Our angle has three features:
- *Two organizing ideas* carried through the field, rather than a catalogue
- A *compact model* at the center of each mechanism, with its first-order condition and comparative static
- The *experiments as the measurement* of the model's parameters, collected in one place

Congiu and Moscati (2022) survey the nudge as a general construct; we instead model the specific environmental internality a nudge is meant to correct.

# Why the environment is different

- Environmental outcomes are the *aggregate* of many individual choices: whether to insulate an attic, which car to buy, how much to fish, what a distant wetland is worth
- Those choices turn on future costs, on other people's welfare, and on one's own self-image, exactly the things the textbook consumer does not register
- So the rational, self-regarding agent describes environmental behavior only in part, and the gaps are large enough to change what policy should do

# The standard benchmark

- Each agent maximizes a stable, self-regarding utility function and discounts the future at a constant rate
- Beliefs are correct, and prices and product attributes are perceived without error
- Under these assumptions the *only* distortion is the externality, the cost one agent's use imposes on others

Everything in the course is a departure from one of these assumptions, introduced one at a time.

# The classical prescription

- An agent who uses quantity $q$, with private benefit $b(q)$ and price $p$, privately chooses $b'(q) = p$
- Each unit also does external damage $\phi$ that she ignores, so efficiency requires $b'(q) = p + \phi$
- A Pigouvian tax equal to that damage closes the gap between the private and the efficient condition

$$t^* = \phi \qquad (\phi = \text{marginal external damage})$$

This remains the point of departure, not the target of the critique: the behavioral analysis is added on top of it.

# A worked Pigouvian benchmark

- Suppose burning a unit of energy does a marginal external damage of, say, \$40 per ton of $CO_2$
- Left alone, the agent uses energy up to the point where her private marginal benefit equals the price, $b'(q) = p$
- Efficiency wants $b'(q) = p + \phi$, so she uses too much: the gap is the externality
- A tax $t = \phi$ raises the price she faces to $p + \phi$, and her private optimum now coincides with the efficient one

Behavioral economics will add a *second* term to this tax, for the harm the agent does to *herself*.

# The behavioral addition: a second wedge

- Behavioral economics adds a second gap, this one *inside* the agent: between the choice she makes and the choice that would serve her own experienced welfare
- By analogy with an externality (a cost to others), call it an *internality*: a cost the choice imposes on the agent's own future or considered self
- It can be present even when the price already reflects the full external damage, so correcting the externality alone does not remove it

Allcott, Mullainathan and Taubinsky (2014).

# Externality vs. internality

- *Externality:* a cost the choice imposes on others, the classical Pigouvian problem
- *Internality:* a cost the choice imposes on the agent's own future or considered self
- Environmental policy generally confronts both at once, and the instrument that fixes one need not fix the other

Example: a car's tailpipe emissions harm others (externality), while the same buyer underweighting future fuel costs harms herself (internality).

# Two organizing ideas

- Some departures *enlarge what the agent values*: she also cares about others, about a moral ideal, or about outcomes relative to a reference point
- There the agent makes no mistake against herself; her welfare is simply richer, so policy works *with* the wider preferences
- Other departures open an *own-welfare wedge*, an internality: present bias and inattention
- There policy can raise the agent's own welfare, and the corrective-tax logic applies

Behavioral welfare economics is what decides which of the two groups a given departure belongs to.

# The two ideas, for policy

- If the departure *enlarges preferences*: design with them, through provision rules, norm-based feedback, the design of an elicitation, or the allocation of a property right
- If the departure opens an *own-welfare wedge*: there are two distortions, so in general two instruments are needed
- Correct the marginal externality and the marginal internality jointly, one instrument aimed at each
- Either way, the size of the behavioral parameter must be *measured*, not assumed

Six departures, and how the survey is organized

# Six departures relevant to the environment

- *Enlarged-preference departures:* other-regarding and warm-glow motives; moral motivation and socially contingent norms; reference dependence and loss aversion
- *Own-welfare-wedge departures:* present bias, and inattention or misperception
- Behavioral welfare analysis then ties them together and asks how to design corrective policy

We take each in turn: first the model, then the laboratory and field experiments that measure its parameter.

# A map (1): departure and parameter

- Every mechanism turns on one parameter, the number policy has to know:
- Warm glow / other-regarding: the warm-glow weight and the crowding-out rate
- Moral motivation / norms: the self-image weight and the strength of the norm
- Reference dependence: the loss-aversion coefficient $\lambda$
- Present bias $\beta$; inattention, the attention weight $m$; internalities, the marginal internality $\gamma$

# A map (2): model, evidence, instrument

- For each departure we pair a compact model, the experiment that measures its parameter, and the instrument it calls for:
- Models: Andreoni; Brekke et al.; Kőszegi–Rabin; Heutel; Gabaix; Farhi–Gabaix
- Evidence: public-goods and common-pool-resource games; the WTA/WTP gap; the energy paradox; the lightbulb experiments
- Instruments: provision policy, norm feedback, valuation design, standards, labels, and the corrective tax

# The six departures at a glance

| Departure | Parameter | Model | Instrument |
|---|---|---|---|
| Warm glow / other-regarding | warm-glow weight | Andreoni; Kotchen | provision policy; certification |
| Moral motivation / norms | self-image weight | Brekke et al.; Nyborg et al. | norm feedback; institutions |
| Reference dependence | $\lambda$ | Kőszegi–Rabin | elicitation design; permit allocation |
| Present bias | $\beta$ | O'Donoghue–Rabin; Heutel | standard or subsidy |
| Inattention | $m$ | Gabaix; Sallee | labels, disclosure, defaults |
| Internalities (welfare) | $\gamma$ | Allcott et al.; Farhi–Gabaix | tax $t^* = \phi + \gamma$ |

Sections 3–8 take these in turn; each has a model and then its experiments.

# Scope and omissions

- We keep a compact model at the center, so several active topics get less room:
- The rebound effect; payments for ecosystem services and the crowding-out they can cause; non-point-source pollution
- The acceptability and perceived fairness of carbon pricing; the growing body of field experiments outside high-income countries
- Behavioral spillovers, a green act in one domain licensing less green behavior in another, only in passing

The cut is model-centered, not a judgment about importance; the surveys cited are the better guide to that breadth.

# Each topic: theory, then its experiments

- For every mechanism we do the model first, then the studies that measure its parameter
- Reading the two together shows where the evidence is thick and where it is thin
- It also keeps the theory honest: a prediction is only as useful as the number attached to it

# Provision of environmental public goods

# Environmental quality as a public good

- Clean air, a stable climate, a healthy fishery: each is a public good, non-rival and hard to exclude
- Hardin's (1968) "tragedy of the commons": self-regarding agents free-ride, so voluntary provision falls short of the efficient level
- The shortfall worsens as the population grows, because each person's own benefit is a smaller share of the total

# Free-riding vs. the efficient level

- Private optimum: $\partial u_i / \partial G = \partial u_i / \partial x_i$, each agent counts only *her own* marginal benefit
- Efficient (Samuelson): $\sum_{j=1}^{n} \partial u_j / \partial G = \partial u_i / \partial x_i$, the *sum* of all agents' marginal benefits
- The private agent internalizes one share and ignores the other $n - 1$, so provision is too low

The gap between these two conditions is exactly the free-riding problem.

# Free-riding: a numeric feel

- Take $n$ identical agents who each value the public good $G$
- Each contributes until *her own* marginal benefit equals marginal cost, while efficiency wants the *sum* of $n$ marginal benefits to equal marginal cost
- So the private level is far below the efficient one, and it is off by roughly a factor of $n$

Large groups do not automatically provide public goods, which is exactly why voluntary provision alone rarely suffices for the environment.

# Large populations

- As $n$ grows, both the fraction of agents who contribute and the amount each gives fall
- So the free-rider problem worsens with the size of the group
- But aggregate voluntary provision need not vanish entirely; it can approach a positive limit

Andreoni (1988).

# A model of private provision

- $n$ agents; each chooses private consumption $x_i$ and a voluntary contribution $g_i \geq 0$
- The public good is the total $G = g_0 + \sum_j g_j$; the budget is $x_i = w_i - \tau_i - g_i$
- The government provides $g_0$, financed by lump-sum taxes $\tau_i$ on the contributors themselves

Each agent takes the others' contributions as given and picks her own: a Nash equilibrium in contributions.

# Pure altruism: neutrality

- In the pure-altruism benchmark utility is $u_i(x_i, G)$: the agent cares about the *total* public good, not about who provides it
- A dollar of government provision is then a perfect substitute for a dollar of her own giving
- So every contributor cuts her own gift one-for-one when the state gives more

$$\frac{dg_i}{dg_0} = -1$$

Neutrality: swapping public for private provision leaves the total $G$ unchanged.

Bergstrom, Blume and Varian (1986).

# Neutrality: the derivation

- In $u_i(x_i, G)$ the agent's desired *total* $G$ depends only on her "full income" $m_i = w_i - \tau_i + g_0 + G_{-i}$
- Raise government provision by $dg_0$, financed by a tax $d\tau_i = dg_0$ on the contributor
- Then $m_i$ is unchanged, so the desired total $G$ is unchanged; with $G_{-i}$ fixed, $g_i$ must fall one-for-one:

$$\frac{dg_i}{dg_0} = -1$$

Bergstrom, Blume and Varian (1986); Appendix A.1.

# Warm glow

- Suppose the *act* of giving carries its own private return, a "warm glow," beyond the total public good it helps produce
- Then the contribution $g_i$ enters utility on its own, alongside consumption $x_i$ and the total $G$

$$U_i = u_i(x_i, G, g_i), \qquad \partial u_i / \partial g_i > 0$$

The agent now gives partly for the private satisfaction of giving, a motive the state cannot supply for her.

Andreoni (1990).

# Warm glow breaks neutrality

- The warm-glow motive does not weaken when the state provides more, so the agent scales her gift back by *less* than one-for-one

$$-1 < \frac{dg_i}{dg_0} < 0$$

- Public and private provision are now imperfect substitutes, and neutrality fails
- How large is the offset? It depends on the warm-glow weight, an empirical object to be measured

# Warm glow: the partial-offset derivation

- The interior condition is now $\partial u_i/\partial G + \partial u_i/\partial g_i = \partial u_i/\partial x_i$
- A tax-financed rise in $g_0$ leaves full income unchanged and raises $G$, lowering the altruism term but leaving the warm-glow term in place
- Cutting $g_i$ one-for-one would over-satisfy the condition, so the agent cuts by *less*: $-1 < dg_i/dg_0 < 0$

The offset approaches $-1$ as the warm-glow weight vanishes and 0 as it dominates. (Appendix A.2)

# Green markets

- A "green" good bundles ordinary consumption with a contribution to environmental quality, providing the public good through the market
- A dollar spent on the green good supplies $a_g$ of the public characteristic; a direct donation supplies $a_d$
- Moving a dollar from a donation to the green good changes provision by $a_g - a_d$: it helps only if $a_g > a_d$
- So a green market can *lower* total provision if it bundles a big private benefit with a small public one; certifying $a_g$ is what tells the two apart

Kotchen (2006).

# Eco-labels as the mechanism

- Because the sign of a green market's effect turns on $a_g$, certification is what makes that public characteristic per dollar visible
- It lets consumers point their warm glow at the products that supply the most environmental quality
- And it lets a regulator tell a green market that helps from one that mostly sells the *feeling* of helping

Eco-labels are, on this reading, the mechanism that decides whether a green market raises provision at all, not a cosmetic addition.

# Provision: the experiments

- The warm-glow weight and the crowding-out rate cannot be read from market data without strong assumptions
- So laboratory and field experiments are the main source of estimates
- We look at both: controlled lab games first, then studies that put real money on real environmental goods

List and Price (2016).

# Public-goods games

- Subjects split an endowment between a private account and a group account that pays everyone
- Contributions start high, at 40 to 60 percent of the endowment, well above the zero the free-riding benchmark predicts
- They decline with repetition but do not fall to zero
- The pattern: *conditional cooperators*, who give more the more others give, mixed with free-riders

Fischbacher et al. (2001); Fehr and Schmidt (1999).

# Why do contributions start high?

- Fehr–Schmidt (1999): many agents dislike *inequitable* outcomes, disliking both falling behind and getting ahead
- This inequity aversion rationalizes the initial cooperation *and* its fragility, as conditional cooperators react to the free-riders they meet
- It is the behavioral basis for the warm-glow and moral-motivation models that follow

Fehr and Schmidt (1999); Fischbacher et al. (2001).

# Fair rules sustain provision

- Kesternich et al. (2014): a public-goods experiment with a *climate* framing and heterogeneous contributors
- When the group can agree on a burden-sharing rule, fairness-based rules sustain higher voluntary provision
- So the *rule*, not only preferences, shapes how much a group provides

Kesternich, Lange and Sturm (2014).

# Punishment sustains cooperation

- Let subjects pay a cost to punish free-riders, and contributions stay high instead of unraveling
- So the *institution* around a commons, not only the preferences of its users, decides whether cooperation survives

This is why monitoring and sanctions matter so much in the field, a theme we return to with Ostrom.

Fehr and Gächter (2000).

# Real-stakes provision experiments

- With real money at stake, voluntary provision is genuine but thin and concentrated in a minority:
- Menges et al. (2005): separate altruism from warm glow in green-electricity donations
- Diederich and Goeschl (2014, 2017): real $CO_2$ contributions, but largely price-inelastic
- Kesternich et al. (2014): fair burden-sharing rules sustain higher provision
- Löschel et al. (2013): buy and retire allowances; positive mean, near-zero median

# Do lab preferences predict field behavior?

- Do the preferences measured in the lab carry to real behavior? Two field studies say yes:
- Fehr and Leibbrandt (2011): fishers who cooperate more and are more patient in the lab exploit their common grounds less in daily life
- Rustagi et al. (2010): Ethiopian forest groups with more conditional cooperators manage their commons better
- So the lab measures have external validity for the settings the models describe

# Moral motivation and socially contingent norms

# Moral motivation

- A second idea: the private return to giving depends on what the agent believes she *ought* to do
- She holds a view of the morally ideal contribution $g_i^*$, and suffers a self-image loss when her own gift falls short of it
- The ideal is a Kantian standard: $g_i^*$ is the contribution that, if everyone chose it, would maximize welfare

Brekke, Kverndokk and Nyborg (2003).

# Self-image utility

- The squared shortfall $(g_i^* - g_i)^2$ is the self-image loss, weighted by $\mu > 0$:

$$U_i = u(x_i, G) - \frac{\mu}{2}(g_i^* - g_i)^2$$

The larger $\mu$, the harder the agent works to close the gap, so she contributes more than a purely self-regarding or altruistic agent would.

# Crowding out of moral motivation

- More public provision lowers the moral ideal $g_i^*$
- The self-image term carries less weight; voluntary giving declines
- A mandate can reduce moral motivation, a genuine welfare cost

$$\frac{dg_i}{dg_0} = \frac{A + \mu\, g_i^{*\prime}}{\mu - A} \in (-1, 0), \qquad A \equiv u_{GG} - 2u_{Gx} + u_{xx} < 0$$

Reading: the offset is negative; at $\mu = 0$ it is the $-1$ of pure altruism, and a positive self-image weight makes it partial.

# Crowding out: the derivation

- Finance the policy with a tax on the contributor, so $dx_i = -(1 + g')\, dg_0$ and $dG = (1 + g')\, dg_0$, with $g' \equiv dg_i / dg_0$
- Differentiate the first-order condition and collect the second-order term $A \equiv u_{GG} - 2u_{Gx} + u_{xx} < 0$:

$$(1 + g')A + \mu(g_i^{*\prime} - g') = 0 \;\Rightarrow\; g' = \frac{A + \mu\, g_i^{*\prime}}{\mu - A}$$

At $\mu = 0$ this is $-1$; for $\mu > 0$ it lies in $(-1, 0)$. (Appendix A.7)

# Socially contingent norms

- Now let the moral return depend on *what others do*: being green is more rewarding, or less costly to self-image, the more others are green
- Write $\theta$ for the green share and $B(\theta)$ for the net private benefit of being green, increasing in $\theta$ (a strategic complementarity)
- Agents differ in cost, so the share who adopt is $F(B(\theta))$, and equilibrium is a fixed point:

$$\theta^* = F\left(B(\theta^*)\right)$$

Nyborg, Howarth and Brekke (2006).

# Multiple equilibria

- Because $B$ is increasing, the fixed-point map can cross the 45-degree line more than once
- So there can be a *low*-participation equilibrium (few act, the norm is weak) and a *high*-participation one (many act, the norm is strong)
- A fixed point is stable where the map is flatter than 45 degrees, $F'(B)\,B' < 1$; an unstable point between them separates their basins

Intuition: beliefs about others are self-confirming, so pessimism sustains low participation and optimism sustains high.

# The fixed-point picture

- Plot the best-response map $\Phi(\theta) = F(B(\theta))$ against the 45-degree line
- Because $B$ rises with $\theta$, the map can cross the 45-degree line more than once
- Stable crossings are where $\Phi$ is *flatter* than 45 degrees; the steeper crossing between them is unstable

Start above the unstable point and participation is drawn up to the high equilibrium; start below it and it slides down to the low one.

# A temporary policy, a permanent effect

- A one-time subsidy or information campaign that lifts participation *past the unstable interior fixed point* tips the economy into the high-participation basin
- Once there, participation keeps rising to the high equilibrium under the ordinary dynamics, and stays after the intervention ends
- So a *temporary* instrument can have a *permanent* effect, unlike the standard model where behavior reverts once the incentive is withdrawn

Nyborg et al. (2016).

# Permanence: the derivation

- Let participation follow $\theta_{t+1} = \Phi(\theta_t)$, with stable points $\theta_L < \theta_H$ and an unstable $\theta_u$ between them
- For $\theta \in (\theta_u, \theta_H)$ the map satisfies $\Phi(\theta) > \theta$, so the sequence $\theta_t$ rises monotonically to $\theta_H$
- A one-period policy that lifts $\theta$ above $\theta_u$ therefore converges to $\theta_H$ under the *original*, unchanged map

A temporary push has a permanent effect. (Appendix A.6)

# An alternative route: social approval

- Rege (2004): the esteem a contributor earns rises with the *number* of contributors
- This produces multiple equilibria by the same strategic complementarity as the contingent-norm model
- So the multiplicity is robust to exactly how the social motive is modeled

Rege (2004); Nyborg et al. (2016).

# Norms and institutions

- A norm is a *state of the population* that institutions can stabilize or erode, not a fixed feature of preferences
- Ostrom's durable commons share design features: clear boundaries, monitoring, graduated sanctions, and collective choice by the users themselves
- These hold participation $\theta$ in the cooperative basin: monitoring keeps beliefs about others accurate, sanctions raise the cost of defecting

Ostrom (1990, 2000).

# Ostrom's design principles

- Durable commons arrangements share features that hold cooperation in place:
- Clear boundaries; monitoring; graduated sanctions; and collective choice by the users themselves
- Ostrom reads these as evolved norms of conditional cooperation, not the outcome of purely self-interested play

Through the norm model, these institutions are what keep $\theta$ in the basin of the cooperative equilibrium.

# Image motivation

# Image motivation

- A third source of the return to giving: the wish to be *seen* as generous
- The payoff has an intrinsic part ($v_a$), an extrinsic part from the reward $y$, and a reputational part $R$ that depends on what the act reveals
- Because observers cannot read motives, a visible reward makes the act an ambiguous signal (generosity or greed) and dilutes its reputational return
- So when reputation matters, paying for a good deed can *reduce* its supply; conspicuous conservation, by contrast, can harness esteem

$$U = (v_a + v_m y)\, a - C(a) + R(E[v_a \mid a, y],\, E[v_m \mid a, y])$$

Bénabou and Tirole (2006, 2011).

# Identity as an asset

- Bénabou–Tirole (2011) extend image motivation to *identity*: beliefs about one's own morality are an asset people protect
- So people may avoid incentives or information that would recast a green act as self-interested
- This strengthens the prediction that visible rewards can backfire, and that some prefer not to know the price of virtue

Bénabou and Tirole (2011).

# Commons experiments

- The common-pool-resource game is the extraction counterpart: subjects draw from a shared resource whose yield falls as total extraction rises
- Extraction exceeds the social optimum but falls well short of the open-access Nash prediction
- Non-binding, face-to-face communication raises cooperation substantially, echoing Ostrom's field findings

Ostrom (1990); Cardenas et al. (2000).

# Institutional crowding-out

- Cardenas et al. add a policy treatment: a modest, externally enforced extraction quota with imperfect monitoring
- It produces *less* cooperation than simply letting the group communicate
- The external rule shifts the decision from a moral to a regulatory frame, crowding out the motivation that had sustained restraint

Cardenas et al. (2000).

# Motivation crowding

- Gneezy–Rustichini: a fine for late daycare pickup *raised* lateness, because "a fine is a price" that replaced a moral obligation
- Frey–Oberholzer-Gee: offering compensation *lowered* residents' acceptance of a local noxious facility, displacing a sense of civic duty
- Bowles: explicit incentives can erode the moral sentiments on which cooperation partly rests

Lesson: a payment or a fee can reduce intrinsic motivation, so the net effect on behavior is an empirical question.

Gneezy and Rustichini (2000); Frey and Oberholzer-Gee (1997); Bowles (2008).

# Home energy reports

- Social-comparison reports cut electricity use by about two percent, and Ayres et al. (2013) replicate this in two further large RCTs (one to two percent)
- Use falls sharply after each report, partly rebounds between reports, then settles as a habit forms
- When the reports stop, the accumulated effect decays only slowly, at 10 to 20 percent a year, evidence of habit formation

This persistence is the empirical counterpart of the "temporary policy, permanent effect" logic.

Allcott (2011); Allcott and Rogers (2014).

# Welfare and the norm channel

- Counting the moral and attention costs the reports impose on recipients, the net welfare gain is positive but well below the value of the energy saved
- Isolating the pure social-comparison channel (reports with no monetary content) shows it delivers much of the effect on its own

Allcott and Kessler (2019); Myers and Souza (2020).

# Water: the heaviest users respond

- Ferraro–Price (2013): norm-based messages to more than 100,000 households cut residential water use
- The largest response comes from the *heaviest* users, and the effect weakens over time
- So a norm nudge is most cost-effective when it is targeted at the top of the distribution

Ferraro and Price (2013).

# Ideology shapes the response

- Costa–Kahn (2013): the electricity response to social comparison is concentrated among self-identified liberals, and is weak or even reverses among conservatives
- Gromet et al. (2013): labeling a product "environmentally friendly" can *lower* its appeal to conservative consumers
- So the framing of a green attribute can defeat the policy in part of the population

Costa and Kahn (2013); Gromet et al. (2013).

# Effects don't transport automatically

- Andor et al. (2020): a large German home-energy-report trial finds savings well below the US estimates
- The same instrument can have quite different effects across countries and institutions
- A caution against importing an effect size from one setting to another

Andor, Gerster, Peters and Schmidt (2020).

# The design of the message

- Telling households how they compare to the average creates a *boomerang*: those below average drift up toward it (Schultz et al.)
- Adding an *injunctive* signal (approval or disapproval) removes the boomerang, and the two norms reinforce each other when aligned (Bonan et al.)
- Even a plain descriptive norm ("most guests reused their towels") beats a conventional green appeal (Goldstein et al.)

So exactly how the norm is conveyed is part of the policy, not a detail of implementation.

Schultz et al. (2007); Bonan et al. (2020); Goldstein et al. (2008).

# Cost, scale, and external validity

- The reports are cost-effective and scale cheaply, but effects at full scale are smaller than in academic trials (DellaVigna–Linos)
- Early-adopting utilities were not representative, so the headline numbers overstate a national rollout (Allcott, site-selection bias)
- Nudges show little crowd-out of other incentives (Brandon et al.); a meta-analysis finds a real but modest, context-dependent effect (Bergquist et al.)

DellaVigna and Linos (2022); Allcott (2015); Brandon et al. (2019); Bergquist et al. (2019).

# From lab to field: a caution

- Levitt–List (2007): laboratory measures of social preferences need not carry to the field, where scrutiny, stakes, and self-selection differ
- Field effects are heterogeneous across populations and decay over time
- So a single average effect is a poor guide to a permanent, universal policy

This caution runs through the whole field, and it is why the same mechanism must be re-measured in each new setting.

# Reference dependence and environmental valuation

# Valuation matters for policy

- So much of policy rests on a valuation number: the damage of a ton of carbon, the worth of a wetland, the compensation owed for a lost amenity
- Standard theory values a change by the equivalent or compensating variation and treats the two as nearly equal for small changes
- Reference dependence predicts a systematic gap between them, and that changes how environmental goods should be valued

# Prospect theory value function

- Outcomes are coded as gains and losses relative to a *reference point* $r$, not as final wealth levels
- The value function is concave over gains, convex over losses, and steeper for losses than for gains
- The reference point is often the status quo, or in Kőszegi–Rabin one's rational expectation of the outcome

Kahneman and Tversky (1979); Tversky and Kahneman (1991).

# Loss aversion

- At the reference point the value function is *kinked*: the loss side is steeper than the gain side by a factor $\lambda$

$$v'_-(0) = \lambda\, v'_+(0), \qquad \lambda > 1$$

So giving up a small amount hurts more than acquiring the same amount pleases. The literature typically estimates $\lambda \approx 2$.

# The endowment effect

- Willingness to accept (WTA) to give up an amenity codes a *loss*, scaled by $\lambda$; willingness to pay (WTP) to obtain it codes a *gain*, not scaled

$$\frac{\text{WTA}}{\text{WTP}} \approx \lambda > 1$$

Example: with $\lambda \approx 2$ a household demands about twice as much to accept the loss of an amenity as it would pay to obtain it, so a benefit-cost number depends on which question is asked.

Horowitz and McConnell (2002).

# WTA/WTP: the derivation

- Value depends on the deviation $z$ from the reference point, kinked at the origin: $v'_-(0) = \lambda\, v'_+(0)$
- WTP buys a *gain*, so it is set by the gain-side slope $v'_+(0)$
- WTA gives up a *loss*, so it is set by the loss-side slope $v'_-(0) = \lambda\, v'_+(0)$

$$\frac{\text{WTA}}{\text{WTP}} = \frac{v'_-(0)}{v'_+(0)} = \lambda$$

Appendix A.4.

# Read with care

- A large WTA/WTP gap does not by itself prove reference dependence
- Hanemann: a standard neoclassical model *already* predicts a gap that widens as the good's substitutes get poorer
- Environmental amenities have poor substitutes, so loss aversion and substitution push the same way and are hard to separate with field data

The gap is *consistent with*, not *proof of*, reference dependence.

Hanemann (1991).

# Probability weighting

- People overweight small probabilities and underweight moderate and large ones (an inverse-S weighting):

$$w(p) = \frac{p^{\eta}}{\left(p^{\eta} + (1-p)^{\eta}\right)^{1/\eta}}, \qquad \eta \approx 0.6$$

- A rare hazard of probability $p$ and damage $L$ is then valued at $w(p)\,L > pL$
- So demand to regulate a salient rare hazard can be strong, while a diffuse risk of similar expected damage is neglected

Tversky and Kahneman (1992).

# Probability weighting and climate tails

- Climate and catastrophic damage are low-probability, high-consequence risks, exactly where weighting and loss aversion bite
- They move valuation away from the expected-utility benchmark, and the *direction* depends on how the risk is described
- So stated valuations of climate risk are very sensitive to framing and to whether the status quo is coded as the reference point

# Policy consequences

# Policy consequences

- Contingent-valuation numbers depend on whether the change is framed as a loss (compensation) or a gain (payment), so the elicitation format is part of the number
- Grandfathered permits are valued as an endowment and resisted more than the standard theory implies, so the initial allocation affects behavior
- Coasean irrelevance (the allocation does not affect efficiency) relies on reference-independent valuations, and here it fails

Kahneman and Knetsch (1992); Diamond and Hausman (1994).

# The contingent-valuation debate

- Kahneman–Knetsch (1992): stated willingness to pay partly buys the *moral satisfaction* of contributing, not the environmental good itself
- Diamond–Hausman (1994): question whether contingent-valuation numbers are informative at all
- Reference dependence *intensifies* this long-running debate rather than settling it: the elicitation format is part of the number

Kahneman and Knetsch (1992); Diamond and Hausman (1994).

# Endowment-effect experiments

- Subjects randomly given an object demand far more to part with it than others will pay to acquire it, even with real goods and repeated trades
- The gap is largest for goods without close substitutes, the category most environmental amenities fall into
- But it shrinks under careful elicitation (Plott–Zeiler) and with market experience (List), so part may be procedure, not preference

Kahneman, Knetsch and Thaler (1990); Plott and Zeiler (2005); List (2003).

# Procedure, or preference?

- Plott–Zeiler (2005): the WTA/WTP gap shrinks or disappears once the elicitation controls for subject misconceptions
- List (2003): market experience attenuates it, so part of the measured gap is procedure, not a stable preference
- Ericson–Fuster (2011): expectations, not only the status quo, set the reference point, the disciplined Kőszegi–Rabin form

The lesson for contingent valuation: a large reported gap has not, by itself, identified its cause.

# Do stated valuations track real ones?

# Do stated valuations track real ones?

- Stated-preference surveys are the main tool for pricing non-market damage, so it matters whether stated values track real ones:
- A *hypothetical bias*: stated values exceed real ones by roughly a third, and by more for willingness to accept than willingness to pay
- A follow-up certainty calibration removes the bias; "cheap talk" scripts do not
- Bulte et al. (2005): stated WTP also shifts with how the cause of the harm is framed

Champ et al. (1997); List and Gallet (2001); Murphy et al. (2005); Blumenschein et al. (2008).

# Evidence on valuing non-market goods

| Study | Design | Result |
|---|---|---|
| Kahneman et al. (1990) | Endowment effect | WTA far exceeds WTP for the same good |
| Horowitz–McConnell (2002) | Review of WTA/WTP | Ratio well above one, highest for non-market goods |
| Plott–Zeiler (2005) | Controlled procedures | Gap shrinks when misconceptions are controlled |
| List (2003) | Market experience | Gap attenuates with trading experience |
| List–Gallet; Murphy et al. | Hypothetical vs. real | Stated exceed real, more so for WTA |
| Blumenschein et al. (2008) | Certainty calibration | Follow-up certainty removes the bias |

# Expectations set the reference point

- The reference point need not be the status quo: *expectations* about the outcome shape what counts as a gain or a loss
- This is the disciplined form the Kőszegi–Rabin account requires, where the reference point is the agent's rational expectation
- It turns loss aversion into a testable feature of preferences rather than a free parameter

Ericson and Fuster (2011).

# Present bias and the energy-efficiency gap

# The energy-efficiency gap

- Households and firms appear to pass up energy-efficiency investments that pay for themselves at plausible discount rates
- Many explanations compete: hidden costs, credit constraints, product quality
- The behavioral explanation: agents discount the near future too steeply and so undervalue the stream of future savings

Jaffe and Stavins (1994); Allcott (2016); Frederick et al. (2002).

# A present-biased investment

- Pay an extra $K$ now for a durable that then yields an energy saving $s$ in each period
- Quasi-hyperbolic discounting adds a factor $\beta \leq 1$ on *all* future periods:

$$U_0 = u_0 + \beta \sum_{t=1}^{T} \delta^t u_t, \qquad \beta \leq 1$$

- True discounted savings are $S = \sum_{t=1}^{T} \delta^t s$

$\beta$ discounts everything beyond today by an extra factor, on top of the usual $\delta^t$; $\beta = 1$ is the standard exponential consumer.

Laibson (1997); O'Donoghue and Rabin (1999).

# The gap

- The present-biased consumer invests only if $\beta S \geq K$, but her own long-run welfare calls for $S \geq K$

$$K \in (\beta S,\ S)$$

- For any cost in this range the efficient investment is worthwhile yet is passed over
- The internality is $(1-\beta)S$: a cost the present self imposes on the future self

Why? The cost $K$ is paid now and the savings come later, so present bias falls entirely on the benefit side.

# Two targets, two instruments

- The externality and the internality are two separate distortions, so in general they need two instruments
- An energy tax $\tau^E = \phi$ corrects the externality
- A subsidy (or an equivalent standard) $\sigma^* = (1 - \beta)S$ corrects the internality
- One instrument cannot do both: a tax at marginal damage leaves the internality; a subsidy at the internality leaves the externality

Heutel (2015).

# Heutel: aligning the biased rule

- Set the energy tax to $\tau^E = \phi$ and the subsidy to the efficient good to $\sigma^* = (1-\beta)S$
- The present-biased agent now invests iff $\beta S + \sigma^* \geq K$
- Substituting $\sigma^* = (1-\beta)S$ turns this into $S \geq K$, exactly the long-run criterion

So the subsidy makes the present-biased agent act as her own long-run self would; a minimum-efficiency standard over the same cost range achieves the same allocation.

# Naive vs. sophisticated

# Naive vs. sophisticated

- A *sophisticated* agent knows her future selves will be present-biased and plans for it, seeking a commitment; a *naif* believes she will act patiently and so procrastinates
- The naif keeps deferring an investment she agrees is worthwhile, and will not seek a commitment on her own
- So a standard that removes the deferral can raise a naif's *own* welfare, where a voluntary instrument would not reach her

The case for a standard over a subsidy is stronger the more consumers are naive, itself a parameter to measure.

O'Donoghue and Rabin (1999).

# Commitment devices for the naive

- A household that agrees an investment is worthwhile yet keeps deferring can be helped by removing the deferral:
- On-bill financing that folds the cost into a familiar payment stream
- Automatic enrollment in an efficiency program, with the freedom to leave
- A contract that installs the efficient option *unless* the household declines

The sophisticated agent would seek these on her own; the naive one will not, which is why offering them as the *default* reaches the households that need them most.

# How big is the gap?

# How big is the gap?

- The undervaluation is real but modest and contested
- Much of the apparent gap dissolves once hidden costs, product quality, and other attributes are accounted for (Allcott–Wozny; Allcott–Greenstone; Gerarden et al.)
- Fowlie et al.: realized savings from a weatherization program fell well short of engineering projections, so the forgone investments were not clearly profitable

This is why the welfare analysis turns on *measuring* the internality, not presuming it.

Allcott and Wozny (2014); Allcott and Greenstone (2012); Gerarden et al. (2017); Fowlie et al. (2018).

# Fowlie et al.: weatherization

- A large randomized weatherization experiment measured *realized* energy savings, not engineering projections
- Realized savings were about *half* the projection, and the upfront cost about *twice* the savings
- That is an average return near $-8\%$ a year: the forgone investments were not clearly profitable after all

A caution against reading present bias into every unexploited investment. (Fowlie et al., 2018)

# Firms forgo efficiency too

- The gap is not confined to households: firms also appear to pass up profitable efficiency investments
- A profit-maximizing firm should not be present-biased the way a person is
- But the managers who decide face split incentives, short horizons, and attention constraints of their own

The organizational version of the gap has its own behavioral and agency roots (Jaffe and Stavins, 1994).

# Eliciting present bias and time preference

# Eliciting present bias and time preference

- Incentivized choices between paid "sooner" and "later" amounts let the analyst back out a discount factor and a present-bias factor:
- Andreoni–Sprenger (2012): convex time budgets, a workhorse elicitation
- Elicited present bias predicts real energy choices: lighting, appliance adoption, and the level of electricity use
- Present-biased households consume about a tenth more electricity

Harrison et al. (2002); Bradford et al. (2017); Schleich et al. (2019); Werthschulte and Löschel (2021); Newell and Siikamäki (2015).

# Elicited time preference and energy choices

| Study | Method | Finding |
|---|---|---|
| Harrison et al. (2002) | Field elicitation | Individual discount rates recovered with real rewards |
| Andreoni–Sprenger (2012) | Convex time budgets | Workhorse method; little present bias once payment risk is handled |
| Newell–Siikamäki (2015) | Elicited rates | Predict energy-efficiency investment |
| Bradford et al. (2017) | Incentivized | Present bias predicts vehicle and lighting choices |
| Werthschulte–Löschel (2021) | Elicit + survey | Present-biased households use $\sim 10\%$ more electricity |

# Climate: a different discounting problem

- Be careful not to conflate two different things:
- *Individual* present bias is an internality to be corrected, an error against one's own future self
- The *social* discount rate is a normative parameter to be chosen; Weitzman shows that uncertainty about it lowers the effective long-run rate, raising the weight on distant damages

Gerlagh and Liski (2018); Weitzman (2001).

# Inattention, misperception, and salience

# Inattention: an alternative to present bias

- A consumer may undervalue future energy costs not because she discounts them, but because she does not *attend* to them, or perceives them with error
- The resulting behavior can look identical to present bias in market data
- But it calls for a different diagnosis and, sometimes, a different remedy (a label rather than a commitment device)

# A sparsity model of attention

- The consumer perceives the energy cost as a weighted average of its true value $e$ and a default anchor $\bar{e}$:

$$\widehat{e} = m\, e + (1 - m)\bar{e}, \qquad m \in [0, 1]$$

The attention weight $m$ says how much she really registers: $m = 1$ is full attention, $m = 0$ means she just uses the anchor. Gabaix (2014).

# The attention trade-off

- Attending more makes the perception $\widehat{e}$ more accurate but is cognitively costly
- So the consumer picks $m$ to trade a quadratic misperception loss against a linear attention cost:

$$\min_{m \in [0,1]} \ \frac{1}{2}(1-m)^2 D + \kappa m$$

- $D$ is the stakes (how much the operating cost matters and varies across products); $\kappa$ is the marginal cost of attention

The next slide solves this for $m^*$.

# Choosing how much attention to pay

- Attention is chosen: trade the misperception loss (stakes $D$) against a cost of attending, $\kappa$ per unit:

$$m^* = \max\left\{0,\ 1 - \frac{\kappa}{D}\right\}$$

Attention rises with the stakes $D$ and falls with the cost $\kappa$; for any positive cost, $m < 1$. Gabaix (2014).

# Underattention distorts choice

- A consumer who trades a higher price $p(q)$ against operating cost sets $p'(q) = -m\,e'(q)$, not the correct $-e'(q)$
- So underattention understates the operating-cost saving by the factor $m$, biasing the choice exactly as a present-bias factor $\beta = m$ would
- That is why $m < 1$ and $\beta < 1$ are so hard to separate in market data: both simply shrink the weight on future costs

# Field evidence on the attention channel

- Real-time in-home displays make households far more responsive to price, by directing attention rather than changing prices (Jessoe–Rapson)
- Prest (2020): under time-of-use pricing, it is *awareness* of the tariff that drives the peak-load response
- Real-time feedback at the moment of use cuts consumption by about a fifth (Tiefenbeck et al.)

Jessoe and Rapson (2014); Tiefenbeck et al. (2018).

# The MPG illusion

- People read fuel economy as if cost were linear in *miles per gallon*, when it is really linear in *gallons per mile*
- So they undervalue the most efficient vehicles: a 10-to-20 MPG upgrade saves more fuel than a 30-to-50 one
- And simply handing buyers fuel-economy information barely changes the vehicles they buy (Allcott–Knittel)

Larrick and Soll (2008); Allcott (2013); Allcott and Knittel (2019).

# Fixing the misperception: metrics and labels

- If the problem is how the attribute is *displayed*, the fix is the metric, not more information:
- Camilleri–Larrick (2014): showing fuel cost over a long distance raises efficient-vehicle choice
- Ungemach et al. (2018): translating an attribute into dollars steers the choice toward it
- Newell–Siikamäki (2014): a monetary lifetime-cost label beats physical-energy information

The specific information a label carries decides whether it corrects the bias.

# Information design does the work

- A disclosure moves behavior only when it lands on the attribute the consumer *underweights*, in the units she uses, at the moment she chooses
- The same fact as an annual figure, a lifetime total, or a running meter meets a different attention weight
- Information supplied in full but pitched at an attribute she already ignores can change nothing

So the *design* of information, not only its provision, is the policy lever.

# Tax salience and shrouding

# Tax salience and shrouding

- Chetty et al.: consumers underreact to a tax added at the register rather than shown in the posted price
- So the *same* tax moves behavior less when it is less salient
- Gabaix–Laibson: firms may deliberately *shroud* an attribute consumers overlook, and competition need not force it back into the open

So an inattention distortion can persist in equilibrium rather than being arbitraged away.

Chetty et al. (2009); Gabaix and Laibson (2006).

# Green defaults

- Because inattentive consumers stick with whatever is chosen for them, making the green option the *default* moves behavior with no price change (Pichert–Katsikopoulos)
- A default matters precisely to the extent that $m < 1$: a fully attentive consumer would just reconsider it
- Coarse labels (ENERGY STAR) help some buyers attend but crowd out others' use of finer cost information (Houde)

Unlike a tax, a default is easy for the attentive consumer to reverse, which is what makes it attractive and also limited.

Thaler and Sunstein (2008); Pichert and Katsikopoulos (2008); Houde (2018).

# Defaults: a diagnostic ambiguity

- A default matters only to the extent that $m < 1$: a fully attentive consumer would reconsider it
- So a default that raises welfare when consumers are *biased* may reduce it when their inattention is a *rational* economy of effort
- It shares the diagnostic ambiguity of the inattention it exploits

That is why the case for a default, like the case for a subsidy or standard, rests on showing the inattention is a genuine bias.

# Defaults move behavior

- Pichert–Katsikopoulos (2008): making the environmentally friendly option the default sharply raises its take-up
- The same works for renewable electricity plans, duplex printing, and thermostat schedules, with no price change at all
- Their power is exactly the inattention they exploit, which is also why their welfare sign needs care

# A caution: rational inattention

# A caution: rational inattention

- If acquiring and processing information is itself costly, then some inattention is privately *optimal*, not an error
- A consumer who rationally declines to compute the lifetime fuel cost of every appliance is not making a mistake policy should fix
- So the underweighting need not be an internality at all: the case for a subsidy or standard has to show a genuine bias

This is the same measurement problem that runs through the whole field.

Sallee (2014).

# Behavioral welfare economics and corrective policy

# The normative problem

- Once choices no longer reveal welfare, a hard question opens up: *whose* preferences count, the deciding self or the experienced self?
- Bernheim–Rangel give choice-theoretic foundations for behavioral welfare economics
- Mullainathan et al. and Chetty develop the reduced-form, pragmatic approach the corrective-tax results below rely on

Bernheim and Rangel (2009); Mullainathan et al. (2012); Chetty (2015).

# The corrective tax

- The tax must close both gaps at once: the marginal externality $\phi$ (harm to others) plus the marginal internality $\gamma$ (harm to oneself)

$$t^* = \phi + \gamma$$

When the consumer undervalues a cost, $\gamma > 0$ and the corrective tax *exceeds* the Pigouvian level; $\gamma = 0$ is the standard rule.

Allcott et al. (2014); Farhi and Gabaix (2020).

# The corrective tax: the derivation

- The consumer chooses $q$ to maximize *decision* utility $v(q) - (p+t)q$, so $v'(q) = p + t$
- Her *experienced* marginal value is $v'(q) - \gamma$ (she overvalues the unit by $\gamma$); consumption also does external damage $\phi$
- Welfare is experienced surplus net of damage, with derivative $[v'(q) - \gamma - p - \phi]\, dq/dt$

set the bracket to zero and use $v'(q) = p + t$ : $\quad t^* = \phi + \gamma$

Appendix A.3.

# Heterogeneity and targeting

- Consumers differ in their internalities, and the bias may be correlated with how much they buy
- The optimal tax weights each consumer's internality by her *responsiveness* to the tax, so those the tax actually moves count most
- If the most biased are the *least* price-responsive, a uniform tax corrects them poorly, and a standard or a targeted nudge can dominate

Farhi and Gabaix (2020); Taubinsky and Rees-Jones (2018).

# The heterogeneous corrective tax

- With heterogeneous internalities, the tax is the externality plus a *responsiveness-weighted* average of the individual internalities:

$$t^* = \phi + \frac{\sum_i \gamma_i \, \partial q_i / \partial t}{\sum_i \partial q_i / \partial t}$$

The consumers who respond most to the tax count most in setting it; $t^* = \phi + \gamma$ is the common-internality case. Farhi and Gabaix (2020).

# The heterogeneous tax: the derivation

- With consumers $i$, each satisfies $v_i'(q_i) = p + t$ and has experienced value $v_i'(q_i) - \gamma_i$
- Welfare's derivative is $\sum_i (t - \gamma_i - \phi)\, \partial q_i / \partial t$; setting it to zero gives:

$$t^* = \phi + \frac{\sum_i \gamma_i \, \partial q_i / \partial t}{\sum_i \partial q_i / \partial t}$$

- The weights are the responsiveness $\partial q_i / \partial t$; a common $\gamma_i = \gamma$ returns $t^* = \phi + \gamma$

Appendix A.5.

# The sufficient statistic for a nudge

- What do we actually need to know to judge a small subsidy or nudge? Only the bias of the consumers it *moves*:

$$\frac{dW}{d\sigma} = \overline{\gamma}^m \frac{dQ}{d\sigma}$$

$\overline{\gamma}^m$ is the average internality among the switchers and $dQ/d\sigma$ the induced change in adoption; the bias of everyone else does not enter.

Allcott and Taubinsky (2015).

# Subsidy vs. ban; the lightbulb verdict

- A *subsidy* raises welfare whenever the marginal consumer (the switcher) is biased
- A *ban* also removes the choice of inframarginal consumers, so it is justified only if even they are biased enough to be better off without the option
- Lightbulbs: the measured bias justifies a moderate subsidy but not a ban

Allcott and Taubinsky (2015).

# Misperceiving the tax schedule

- Rees-Jones–Taubinsky (2020): people asked to forecast the income-tax schedule replace it with their *average* rather than marginal rate
- So they misperceive the very incentive a tax is meant to create
- A corrective tax works through a schedule that consumers do not perceive correctly

Rees-Jones and Taubinsky (2020).

# Under- and overreaction to a hidden tax

- Morrison–Taubinsky (2023): varying a shrouded sales tax within subjects, some *under*react and others *over*react
- Both errors shrink as the stakes rise
- Evidence that attention is the costly, heterogeneous choice the welfare formulas assume

Morrison and Taubinsky (2023).

# Attention varies with income

- Goldin–Homonoff (2013): low-income buyers react to a cigarette tax only when it is in the posted price
- So the attention gap itself varies with income
- This is what makes the distributional analysis of a corrective tax turn on *who attends*

Goldin and Homonoff (2013).

# Which nudge, and for whom

- An *informational* nudge (a label, a lifetime-cost disclosure) corrects a belief; if the internality is real its welfare sign is clear, because it expands the consumer's effective choice set
- A *default* works through the same inattention that caused the problem, so it is harder to defend on welfarist grounds
- Distribution matters: energy taxes are regressive, so return the revenue to offset the incidence, or prefer a standard where households cannot switch

Allcott, Lockwood and Taubinsky (2019).

# Distribution and the fiscal system

- Energy is a necessity, so a corrective energy tax falls harder on low-income households as a share of budget
- If bias is *not* stronger among the poor, taxing their energy to correct an internality they lack is a pure loss to them
- Allcott–Lockwood–Taubinsky: the optimal corrective tax, when bias and price-sensitivity both vary with income, combines the corrective and redistributive motives

In practice: return the revenue to offset regressivity, or prefer a standard where households cannot switch.

Allcott, Lockwood and Taubinsky (2019).

# Measuring the internality directly

- The newest field designs measure the internality directly rather than inferring it:
- Give consumers customized energy-cost information and read the shift in demand as the size of the attention or belief gap (Boogen et al.; Beattie et al.)
- Cross-randomize information and subsidies to estimate structurally the welfare cost of imperfect information (Rodemeier–Löschel)
- Davis–Metcalf: state-specific cost labels improve choices where a generic national label does not

Boogen et al. (2022); Beattie et al. (2022); Rodemeier and Löschel (2025).

# Measuring the behavioral parameters

# Three research designs

- Incentivized elicitation: a clean parameter, far from the decision of interest
- Structural estimation from market data: close to the decision, but cannot separate a wedge from an unmeasured cost
- Randomized field experiment: causal, yet an average effect, and it travels poorly

Trust a parameter where two of the three agree.

# Master estimates (1)

| Parameter | Estimate | Reading |
|---|---|---|
| Loss aversion $\lambda$ | About 2; WTA/WTP $\approx$ 3 private, $\approx$ 10 non-market | Poor substitutes inflate the ratio |
| Present bias $\beta$ | Below 1; biased households use $\sim$ 10% more electricity | Sign clear, size varies |
| Value of future savings | \$1 of future fuel priced near 76 cents | Undervaluation $\approx$ a quarter |
| Realized vs. projected | Realized $\approx$ half; return near $-8\%$ a year | Not clearly profitable |
| Attention weight $m$ | Real-time feedback cuts use $\sim$ 20% | Responds at the point of use |

# Master estimates (2)

| Parameter | Estimate | Reading |
|---|---|---|
| Marginal internality $\gamma$ | Modest for lightbulbs: a subsidy helps, a ban does not | Smaller than paradox claims |
| Norm feedback | $\approx$ 2% cut; decays 10–20% a year; smaller at scale | Context-dependent |
| Crowding-out rate | Sign established; coefficient elusive | Direction clear |
| Hypothetical bias | Stated values exceed real by $\approx$ a third | Larger for WTA |

# Reading the table

- Every parameter varies by close to an order of magnitude across settings
- A single number transported across studies is a poor guide to a policy elsewhere
- Policy needs the bias of the *marginal* consumer, which the evidence least often reports

# Loss aversion: the best measured

- $\lambda$ sits near 2 across many elicitations: giving a good up registers as about twice the change that acquiring it does
- Its environmental face is the WTA/WTP gap, near 3 for ordinary goods and near 10 for non-market goods (Horowitz–McConnell)
- But poor substitutes widen the ratio for reasons unrelated to reference dependence (Hanemann), so it bounds $\lambda$ from above

# Present bias: two methods, one lesson

- Elicitations recover $\beta < 1$ and a long-run discount factor, and both predict energy choices (Bradford et al.; Werthschulte–Löschel)
- Market data put the combined undervaluation more sharply: \$1 of future fuel priced near 76 cents (Allcott–Wozny)
- Whether that quarter is present bias, inattention, or an unseen cost is what market data cannot settle, and why the elicitations matter

# The internality: needed most, measured least

- $\gamma$ is what the welfare analysis turns on, and the hardest to recover, a gap between a choice and an unobserved welfare
- Allcott–Taubinsky estimate it for lightbulbs and find it modest: enough for a subsidy, not for a ban
- Newer field designs recover it by cross-randomizing information and prices (Boogen; Beattie; Rodemeier–Löschel; Davis–Metcalf)

Common finding: real, but smaller than the early energy-paradox claims, and heterogeneous across consumers.

# Nudges, crowding-out, weighting

- Nudge effects are positive, modest, and context-dependent, about 2%, decaying over time and shrinking at scale
- Warm glow and crowding-out are settled in *sign* but elusive in *magnitude*: the crowding-out rate is rarely a single coefficient
- Probability weighting is the least pinned parameter, measured in the lab and only inferred for the environmental risks that matter

# Mechanisms across environmental domains

# The same mechanisms, different weights

- The six mechanisms are general, but their *relative* importance shifts across settings
- What decides the weights is the *structure of the decision*: who chooses, over what horizon, and who bears the cost
- So choosing the right instrument needs domain knowledge, not just a general behavioral principle

# Domains, mechanisms, instruments

| Domain | Dominant mechanism | Suited instrument |
|---|---|---|
| Household energy | present bias, inattention, norms | standard/subsidy; carbon price; reports |
| Transport | inattention to fuel economy | fuel-cost labels; salient fuel tax |
| Residential water | social comparison, norms | feedback, targeted at heavy users |
| Waste, recycling | warm glow, moral motivation | provision design; avoid crowd-out |
| Climate (individual) | probability weighting, $\lambda$ | risk communication; framing |
| Climate (collective) | social discount rate | normative choice, not a corrective tax |
| Local commons | other-regarding, cooperation | boundaries, monitoring, sanctions |

# Household energy and transport

- *Household energy*: an upfront cost against future savings, so present bias ($\beta$) and inattention ($m$) lead, with norms entering through feedback. Instruments: standards, efficiency subsidies, home energy reports
- *Transport*: the same biases at the point of purchase, sharpened by the fuel-economy misperception; a fuel tax must be shown saliently to move driving. Instruments: fuel-cost labels, a salient fuel tax

# Water, waste, and recycling

- *Residential water*: social comparison and norms are the active mechanism, with strong heterogeneity by prior use and ideology. Instrument: norm-based feedback, targeted at heavy users
- *Waste and recycling*: the private material return is tiny, so behavior rests on conscience, self-image, and norms, exactly where a fee can crowd motivation out. Instrument: provision design that protects intrinsic motivation

# Climate and the local commons

- *Climate, individual*: probability weighting and loss aversion shape how citizens value rare catastrophic risk, and norms can drive self-reinforcing shifts. *Climate, collective*: the social discount rate, a normative choice, not a corrective instrument
- *Local commons* (fisheries, grazing, groundwater): other-regarding preferences and conditional cooperation are decisive, sustained by communication and self-management, and eroded by heavy-handed external rules

# Buildings: a split incentive

- The landlord pays $K$; the tenant gets the operating saving $S$
- The landlord recovers only a fraction $\theta$ through rent: installs iff $\theta S \geq K$
- With a present-biased tenant the frictions compound: installs iff $\theta\beta S \geq K$
- A construction standard binds the party that makes the decision

$$K \in (\theta\beta S,\ S) \text{ is forgone}$$

# Agriculture and land use

- A farmer's conservation choice carries both a private return and an external benefit
- Other-regarding motives and social norms operate *alongside* payments for ecosystem services
- Those payments can crowd the motives out, so the caution about a fee for a moral act applies to a payment for a conservation one

Like buildings, this domain adds no new mechanism; the structure of the decision selects which of the six matters.

# Implications for policy design

# Correct both wedges jointly

- The organizing result: the optimal corrective tax is the marginal externality *plus* the marginal internality, two separate objects with separate determinants
- A carbon or energy tax at marginal damage remains the right instrument for the externality; behavioral considerations do not raise or lower it
- What they add is a *second* correction, for the consumer's undervaluation of her own future costs

Neglect the internality and you underdo efficiency policy; rely on nudges alone and you leave the externality uncorrected.

# Prices for the externality, nudges for the internality

- A nudge fits the internality (a property of cognition); a tax fits the externality (a property of the good's use)
- This is comparative advantage, not exclusivity: a nudge also cuts the externality a little, and a salient tax also cuts the internality a little
- But a green nudge is not a full substitute for a price: it is not calibrated to the damage others bear

Ito et al.: moral suasion produced only a short-lived response, while a price change was larger and more persistent.

Carlsson et al. (2021); Ito et al. (2018).

# Target on the margin

- Because the welfare gain from a subsidy or nudge depends on the bias of the consumers it *moves*, evaluate policy against the marginal consumer, not the average
- The same criterion warns against bans, which remove the choices of inframarginal consumers who may not be biased at all
- So the internality has to be measured for the specific population a policy will move, a demanding empirical standard

# Attend to beliefs and acceptability

- The political feasibility of a corrective price turns on what citizens *believe* about it, and those beliefs are often wrong
- Douenne–Fabre: opposition to a French carbon tax rests on overestimated personal costs and misperceived incidence
- Dechezleprêtre et al.: credible information about a policy's effects raises support across 20 countries; Andre et al.: people underestimate others' willingness to act

Acceptability is itself a behavioral object, so the beliefs a policy must overcome are part of its design.

Douenne and Fabre (2022); Dechezleprêtre et al. (2025); Andre et al. (2024).

# Take heterogeneity and framing seriously

- Responses to green interventions differ across ideology and other observable groups
- A green framing can *raise* demand in one group and *lower* it in another
- A national instrument that ignores this will misestimate its own effect, and a framing chosen without regard to the audience can defeat the policy

The design of the message is part of the policy, not a detail of implementation.

# Protect moral motivation; weigh the equilibrium

# Protect moral motivation; weigh the equilibrium

- A monetary instrument can *erode* the intrinsic and social motivation that sustains voluntary restraint, so where that motivation is strong a heavy-handed incentive can backfire
- A subsidy or nudge also changes demand, and firms respond by adjusting prices and the attributes they make salient
- So the welfare effect and incidence of a behavioral policy are general-equilibrium objects, not partial ones

# Combine instruments; regulators are behavioral too

- With two distortions and no single instrument that corrects both, the design question is which *combination* assigns each instrument to the distortion it fits
- Ranking instruments in the abstract (tax vs. standard vs. nudge) asks a question the theory says is ill-posed
- And the same departures describe the officials who design policy: reference dependence in permit allocation, present bias in long-horizon targets, inattention in disclosure design

# Three worked cases

# Household energy: an own-welfare wedge

- A carbon externality $\phi$ plus an internality $\gamma$ in durable purchases
- Carbon price at $\phi$; efficiency standard or subsidy for $\gamma$
- A subsidy passes the marginal-consumer test; a ban does not (lightbulbs)
- A home energy report is a cheap complement; return revenue for the regressivity

# A local fishery: an enlarged preference

- The departure is a preference, not a wedge against the fisher's own welfare
- Cooperation is sustained by conditional cooperation and communication
- A corrective tax is the wrong tool; a weak quota can crowd out restraint
- Support the cooperation: boundaries, monitoring, graduated sanctions

# Transport: a misperceived attribute

- Drivers misread fuel economy (the MPG illusion) and undervalue efficient cars
- Fits a display, not a price: a fuel-cost label in dollars over the vehicle's life
- A fuel tax still corrects the externality, the more so when shown saliently
- Information alone moves choice little: size the instrument to the wedge

# Frontier: separating look-alikes

- Present bias and inattention imply nearly identical market behavior but call for different instruments (a commitment or standard vs. a disclosure or default)
- Market data on the undervaluation cannot tell them apart
- Designs that vary the *salience* of future costs independently of the *timing* of payments would identify the two channels

The same design problem recurs for warm glow vs. altruism, and a genuine reference-dependent preference vs. a manipulable frame.

# Frontier: the marginal wedge, and firms

- Policy needs the average bias among the *switching* consumers, but the evidence reports population averages
- How firms adjust prices, attributes, and shrouding when consumers are biased is understood in theory and rarely measured
- Almost the whole literature sets a biased consumer against a passive, competitive supply side, leaving market power and pass-through open

# Directions for future research

- Separate present bias from inattention by design
- Measure the marginal wedge, not the population average
- General-equilibrium and long-run responses to behavioral policy
- Firmer normative foundations for welfare with endogenous reference points
- Transportability across settings; firms and market structure

# Conclusion

# What the field does not claim

- Behavioral environmental economics *embeds* the Pigouvian analysis rather than overturning it
- The externality remains, and the corrective price is still the right instrument for it
- In the special case of unbiased agents, every result returns to its classical form

The contribution is an *enlargement* of the standard analysis, a second wedge and a second instrument, not a rejection of it.

# The two ideas, restated

- Some departures enlarge preferences; policy works with them
- Others open an internality beside the externality; correct both
- Evaluate against the consumers a policy moves
- The open problems are, in essence, problems of measurement

# Appendix: the derivations at a glance

- A.1 Neutrality: $dg_i/dg_0 = -1$ under pure altruism (full income unchanged)
- A.2 Warm glow: $-1 < dg_i/dg_0 < 0$ (the warm-glow term is untouched)
- A.3 Corrective tax: $t^* = \phi + \gamma$ (decision vs. experienced value)
- A.4 Endowment: WTA/WTP $= \lambda$ (loss-side vs. gain-side slope)
- A.5 Heterogeneous tax: externality plus responsiveness-weighted internality
- A.6 Norm permanence: $\Phi(\theta) > \theta$ above $\theta_u \Rightarrow \theta \to \theta_H$
- A.7 Moral crowding-out: $g' = (A + \mu g_i^{*\prime})/(\mu - A)$

# Thank you

Behavioral Environmental Economics

*Theory and Experiments*

Comments welcome.