Diagnostic Utility:

Theoretical model and experiment

Josep Serrano

Advisor: Han Bleichrodt

20 September 2026

Abstract

This paper translates the qualitative principles of Bodner and Prelec's self-signaling model into explicit assumptions and conditional predictions. It distinguishes the direct consequences of an action from the value of what that action reveals about the chooser. Preferences compare direct acts together with beliefs about disposition; Bayesian updating is stated separately from those preferences. The framework allows a general ranking of self-beliefs and identifies the additional assumptions behind BP's additive, affine specification. A proposed experiment applies the model to paying for relief from a task, either for another person or for oneself. A freedom-of-choice monetary benchmark is combined with prior and posterior questions. The design explains which comparisons could separate diagnostic utility from direct task value, and why beliefs and willingness to pay alone do not identify that contribution without further restrictions.

1 Purpose and scope

The objective is to make Bodner and Prelec's (BP henceforth) qualitative claims mathematically precise enough to evaluate in an experiment. Their model distinguishes the direct consequences of an action from the value of what that action reveals about the person's disposition. Their explicit assumptions combine optimization with interpretations that account for the motivation to self-signal [1, pp. 4–7].

The analysis proceeds in three steps: define the choice objects; state the model's substantive principles; derive restrictions under stated experimental assumptions. The principles are maintained hypotheses, not conclusions of a representation theorem. In particular, additivity, expected utility, and Bayesian interpretation are stated separately so that their implications can be examined separately.

The mathematical language follows Savage's distinction between states, consequences, and acts [2]. A full axiomatization is left for later work. Here, preferences compare an act together with the belief about oneself that the choice produces. The model also needs a rule explaining how that belief is formed.

2 Literature and contribution

2.1 Related literature

Bodner and Prelec [1] provide the starting point: people may value an action both for its consequences and for the information it provides about their own disposition. Their model separates the preference that actually guides behavior from the person's uncertain view of that preference. The present paper follows this mechanism and makes its preference assumptions, interpretation rules and experimental implications explicit.

Related work shows how self-image can influence other economic decisions. Bénabou and Tirole [4] explain personal rules through self-reputation about willpower: a lapse can undermine confidence in future self-control. Their dynamic mechanism differs from the current model's direct valuation of self-beliefs at a single decision. Bénabou and Tirole [5] also show how material incentives can change the interpretation of prosocial actions, potentially weakening motives based on reputation or self-respect. This motivates treating an action's meaning as dependent on its incentives rather than assuming a fixed signal.

Diagnostic utility must also be distinguished from other reasons to bear a cost. Andreoni's warm-glow model [6] allows giving itself to enter utility alongside the benefit produced by giving. In the present framework, such a direct reward need not involve learning about one's disposition. Gächter, Molleman and Nosenzo [7] document costly compliance with arbitrary rules in anonymous settings and examine intrinsic respect for rules and social expectations. Their evidence makes clear why costly rule-following alone would not identify a diagnostic motive.

The experimental approach draws on Feldhaus et al. [3], who compare payment for guaranteed implementation with a monetary lottery benchmark to measure the intrinsic value of choice. The proposed helping experiment adds explicit reports of prior and posterior self-beliefs. Because relief from another person's task and relief from one's own task have different direct values, its identification argument retains that difference rather than assigning the entire corrected premium to self-image.

2.2 Contribution and possible applications

The paper's contribution is to organize BP's principles into a transparent framework connecting preferences, beliefs and observable choices. It separates predictions based on an ordinal ranking of self-beliefs from those requiring additive or affine utility, and distinguishes Bayesian consistency from the stronger requirement that interpretations agree with behavior. The experiment illustrates how those distinctions guide measurement and which restrictions are needed to recover diagnostic value. The contribution is a formal clarification and an experimental proposal; it does not claim a full axiomatization or an established empirical estimate.

One application is *everyday helping*: returning a shopping cart, assisting a colleague or completing a small task for someone else. The action can improve another person's circumstances and support a favorable view of oneself as considerate. Donations and environmentally responsible choices offer related applications, provided that their direct benefits and other motives are accounted for. The present experiment concentrates on a small, concrete instance of helping.

A second application is *truth-telling*. Gneezy, Kajackaite and Sobel [8] study costly honesty and partial lying in a reporting task. Applying the present framework would ask whether an honest report is valued partly because of what it reveals about the reporter, alongside any direct aversion to lying. This would require beliefs conditional on both the information received and the report made; observed honesty alone would not distinguish the motives.

A third application is *self-imposed rules*, such as study targets, spending limits or promises to finish a task. Keeping a rule may support a belief in one's reliability, while breaking it may weaken that belief. Comparing voluntary and assigned rules could examine whether authorship changes their diagnostic meaning. Such an extension would also have to account for selection into voluntary rules and any direct value of keeping a commitment. These are possible uses of the framework, rather than predictions already tested by the helping experiment.

3 States, acts, and preferences

Let Ω be a finite nonempty set of external states, X a nonempty set of direct consequences, Q a finite nonempty set of dispositions, and $\mathcal{A}$ a fixed finite nonempty action set. Finite states and dispositions keep the formulation simple.

An action $a \in \mathcal{A}$ induces an act

$$f_a : \Omega \longrightarrow X.$$

Thus $f_a(\omega)$ is its direct consequence if external state ω occurs. A consequence includes all relevant direct effects, such as payment, harm to others, and any direct moral cost of choosing the action.

Let $\rho : \Omega \to [0,1]$, with $\sum_\omega \rho(\omega) = 1$, describe the decision-maker's subjective beliefs about external states. The same ρ is used across the actions and dispositions considered here. I use subjective probabilities but do not derive them from Savage's axioms. These beliefs concern external uncertainty; they are distinct from beliefs about one's own disposition.

The actual disposition $q \in Q$ describes the psychological traits that affect choice. It is distinct from the external state ω. The reflective self knows the possible dispositions in Q, but does not directly observe q and cannot fully recover it through introspection. A disposition can therefore affect behavior even when the person cannot consciously identify it. We use q for the actual disposition and θ for a possible one.

Write

$$\Delta(Q) = \left\{ \mu : Q \to [0,1] : \sum_{\theta \in Q} \mu(\theta) = 1 \right\}.$$

A prior $F \in \Delta(Q)$ describes the reflective self's beliefs before interpreting the choice: $F(\theta)$ is the probability assigned to having disposition θ. For simplicity, assume $F(\theta) > 0$ for every θ. This full-support assumption means that no disposition in Q is ruled out beforehand. The prior is a primitive of the model; it is not derived from preference axioms.

Each action is associated with a belief-update rule B_a. For a fixed interpretation rule, it takes a prior to a posterior in $\Delta(Q)$:

$$F \xrightarrow{a} B_a(F) = \mu_a. \tag{1}$$

To compare all available actions, the interpretation rule must supply $B_a(F)$ for every $a \in \mathcal{A}$. Under Bayesian interpretation, the formula applies when the action has positive predicted probability; otherwise its posterior needs a separate rule.

The posterior μ_a, also written $\mu_a(\theta) = \mu(\theta \mid a)$, changes the belief about q, while q itself stays fixed. An act therefore implies

$$\widehat{f}_a(\omega) = (f_a(\omega), B_a(F)) = (f_a(\omega), \mu_a). \tag{2}$$

The posterior is formed when the action is chosen. It can serve as a prior in a subsequent decision if no further information arrives; the present analysis concerns a single choice.

For each actual q, define a preference relation $\succeq_q$ on

$$\mathcal{H} = X^\Omega \times \Delta(Q),$$

where X^Ω is the set of all functions from Ω to X. The statement $(f, \mu) \succeq_q (g, \nu)$ means that disposition q weakly prefers act f with posterior μ to act g with posterior ν. This larger domain allows hypothetical comparisons even when some pairs cannot arise from the available actions.

Holding the prior F and the update rules fixed, preferences over actions are induced by this single relation:

$$a \succeq_q^{\mathcal{A}} b \quad \Longleftrightarrow \quad (f_a, B_a(F)) \succeq_q (f_b, B_b(F)). \tag{3}$$

The posterior is part of the complete object being evaluated. It describes a belief about disposition. Defining preferences this way allows comparisons across different posteriors. A separate condition, stated below, requires the interpretation rule to agree with the behavior it predicts.

4 The model's principles

4.1 Principle 1: an action can provide evidence without changing disposition

At the moment of choice, q is fixed across the available actions. The act f_a records what choosing a directly causes; μ_a records the chooser's resulting belief about disposition. Effects such as payment, harm, and a direct cost of dishonesty belong in f_a. The present model leaves changes in disposition over time for later work. This states BP's distinction between causal and diagnostic effects (pp. 1–5 and 7–8).

4.2 Principle 2: direct utility depends on actual disposition

Let $u : X \times Q \to \mathbb{R}$ be direct utility. Expected direct utility is

$$G(a, q) = \sum_{\omega \in \Omega} \rho(\omega) u(f_a(\omega), q). \tag{4}$$

Only external uncertainty is averaged here. The actual q governs direct utility even though the reflective self is unsure what q is. Expected utility is adopted for simplicity. The distinction between actual disposition and beliefs about it follows BP (p. 5).

4.3 Principle 3: preferences over self-beliefs

Ranking dispositions. Start with a complete and transitive ranking $\succeq^Q$ on Q, common across actual dispositions. It describes which disposition the person would prefer to believe they have, with direct consequences held fixed. For example, an honest disposition can be preferred to a dishonest one. This ranking is a preference assumption; it does not follow from the prior or from Bayes' rule.

For each $\theta \in Q$, let $e_\theta \in \Delta(Q)$ assign probability one to θ. A preference $\succeq^D$ over self-beliefs must agree with the disposition ranking at certainty:

$$e_\theta \succeq^D e_\eta \quad \iff \quad \theta \succeq^Q \eta. \tag{5}$$

This condition alone does not rank uncertain self-beliefs.

Preference over distributions. Assume the preferences $\succeq_q$ defined above are complete and transitive. Require a common diagnostic ranking $\succeq^D$ such that, for every direct act f and actual disposition q,

$$\mu \succeq^D \nu \quad \iff \quad (f, \mu) \succeq_q (f, \nu). \tag{6}$$

Thus the ranking of self-beliefs does not depend on the common direct act or on actual disposition. This is an explicit restriction on the joint preference. It gives $\succeq^D$ completeness and transitivity without assigning numerical utilities. Write $\succ^D$ for strict preference and $\sim^D$ for indifference.

To connect uncertain beliefs to the ranking of dispositions, impose *diagnostic monotonicity*: a belief is weakly preferred if it assigns at least as much probability to every group of better dispositions. Formally,

$$\left[\sum_{\eta:\, \eta \succeq^Q \theta} \mu(\eta) \geq \sum_{\eta:\, \eta \succeq^Q \theta} \nu(\eta) \quad \text{for every } \theta \in Q \right] \quad \Longrightarrow \quad \mu \succeq^D \nu. \tag{7}$$

This is a first-order stochastic dominance condition. It is an additional assumption, not a deduction from ranking the dispositions. With three or more distinct ranks, it does not determine every comparison. For example, it does not rank certainty of an intermediate disposition against an equal chance of the best and worst. Those comparisons require further preferences.

The binary case. Let $Q = \{H, L\}$, where H denotes a more valued disposition and L a less valued one, with $H \succ^Q L$. In the helping experiment, these refer to stronger and weaker concern for others' effort. Impose strict preference whenever the probability of H is higher. Since a belief is then determined entirely by $\mu(H)$, this gives

$$\boxed{\mu \succeq^D \nu \quad \Longleftrightarrow \quad \mu(H) \geq \nu(H).} \tag{8}$$

For example, an 80% probability of having the more valued disposition is preferred to a 30% probability, with direct consequences fixed. No numerical value for the disposition is needed. The person still chooses among actions and their associated posteriors, rather than freely choosing which belief to hold.

An additional specification for numerical comparisons. The ranking alone does not say how much money or other direct value the person would sacrifice for a better self-belief. For numerical comparisons, assume a real-valued representation $D : \Delta(Q) \to \mathbb{R}$ of $\succeq^D$, and assume the joint preference has the additive representation

$$J_q(f, \mu) = \sum_{\omega \in \Omega} \rho(\omega) u(f(\omega), q) + D(\mu). \tag{9}$$

These are further assumptions. The common D must respect the diagnostic ranking, but need not be affine. For available actions, write

$$\begin{aligned} U(a, q) &= G(a, q) + d(a), \qquad d(a) = D(B_a(F)), \\ a \succeq_q^{\mathcal{A}} b \quad &\Longleftrightarrow \quad G(a, q) + D(B_a(F)) \geq G(b, q) + D(B_b(F)). \end{aligned}$$

In the binary case, any strictly increasing function $\phi : [0, 1] \to \mathbb{R}$ gives the required diagnostic ranking through $D(\mu) = \phi(\mu(H))$. Its shape and scale relative to direct utility remain to be specified or measured. Although $\mu(H)$ itself is an affine representation of this binary ranking, choosing a nonlinear ϕ changes its tradeoff against direct utility. It does not change the ranking of self-beliefs alone.

BP's affine specification as a special case. BP's expected diagnostic value imposes the stronger condition

$$D(\alpha\mu + (1 - \alpha)\nu) = \alpha D(\mu) + (1 - \alpha) D(\nu), \qquad \alpha \in [0, 1]. \tag{10}$$

Only in this special case do we introduce the type-value function through

$$V(\theta) := D(e_\theta), \qquad D(\mu) = \sum_{\theta \in Q} \mu(\theta) V(\theta). \tag{11}$$

The second equality follows from affinity and $\mu = \sum_\theta \mu(\theta) e_\theta$. Thus $V(\theta)$ is the value of believing with certainty that one's disposition is θ. Together with additivity, this recovers BP's equation (1), pp. 4–5. The ordinal model does not require it. Neither the ordinal conditions nor these numerical specifications are derived here from Savage's axioms.

The later numerical restrictions use the additive specification (9); affinity is required only where stated. Across experimental conditions, stability of the joint preference and of any numerical specification is an additional assumption. Random assignment alone does not establish it.

4.4 Principle 4: choice anticipates its interpretation

A choice rule $\sigma(a \mid q) \geq 0$, with $\sum_a \sigma(a \mid q) = 1$ for every q, gives the probability of each action under each disposition. Optimal choice can be stated directly in terms of preferences:

$$\sigma(a \mid q) > 0 \quad \Longrightarrow \quad (f_a, \mu_a) \succeq_q (f_b, \mu_b) \quad \text{for every } b \in \mathcal{A}. \tag{12}$$

Under the additional additive representation, this is equivalent to

$$\sigma(a \mid q) > 0 \quad \Longrightarrow \quad a \in \arg\max_{b \in \mathcal{A}} \{G(b, q) + D(\mu_b)\}.$$

Thus an action is chosen only if it is best after accounting for its interpretation. A disposition may randomize among equally good actions. The interpretation assigned to each possible action stays fixed during this comparison: choosing b carries μ_b, rather than triggering a new solution of the whole model. This is BP's Optimization assumption (p. 6).

4.5 Principle 5: true interpretations account for the motive to self-signal

Under true interpretations, the reflective self uses the choice rule that includes the desire to self-signal. Its predicted probability of action a is

$$m(a) = \sum_{\theta \in Q} F(\theta)\sigma(a \mid \theta).$$

Holding σ fixed, the Bayesian update operator for a prior $\widetilde{F} \in \Delta(Q)$ is

$$B_a(\widetilde{F})(\theta) = \frac{\widetilde{F}(\theta)\sigma(a \mid \theta)}{\sum_{\eta \in Q} \widetilde{F}(\eta)\sigma(a \mid \eta)}, \tag{13}$$

whenever the denominator is positive. This maps each such prior to a posterior in $\Delta(Q)$. At the model's prior F, for any action with $m(a) > 0$,

$$\mu_a(\theta) = B_a(F)(\theta) = \frac{F(\theta)\sigma(a \mid \theta)}{m(a)}. \tag{14}$$

The notation B_a leaves its dependence on σ implicit. Changing the prior and solving for a new equilibrium may also change σ, and hence the update operator. For a given equilibrium, the operator stays fixed when comparing actions. The prior F is the starting belief; μ_a is the belief after interpreting the action. Bayes' rule weights each prior probability by that disposition's likelihood of choosing the action. Simply listing the dispositions for which the action is optimal would not supply these likelihoods.

An equilibrium is a choice rule σ and a posterior μ_a for every action satisfying both optimality (12) and Bayesian consistency (14) wherever $m(a) > 0$. This states BP's True Interpretations assumption (pp. 6–7). The two conditions must be solved together. Defining them does not prove that a solution always exists or is unique. Including posteriors in preferences makes choices well defined for a given interpretation rule; it does not by itself explain or identify that rule.

BP also considers two alternatives (pp. 5–6). Under *exogenous interpretations*, the posterior for each action is supplied directly. Under *face-value interpretations*, the reflective self applies Bayes' rule using a rule σ^0 that maximizes direct utility alone, although actual choices can include diagnostic utility. These are different versions of the interpretation rule.

For an action with $m(a) = 0$, Bayes' rule supplies no posterior. Such actions are called *off path*. Within the additive specification, one optional restriction, based on BP's plausibility rule (p. 7),

assigns them to dispositions that would lose the least from choosing them. Define

$$\begin{aligned} R(q,a) &= U(a,q) - \max_{b\in\mathcal{A}} U(b,q), \\ m(a) = 0 \text{ and } \mu_a(q) > 0 \quad &\Longrightarrow \quad q \in \arg\max_{\theta\in Q} R(\theta,a). \end{aligned} \tag{15}$$

Ties require a further rule for assigning probabilities. Comparing losses across dispositions also requires the assumed common utility scale; ordinal rankings alone do not supply it. This optional restriction is needed only when it is used to select interpretations of unused actions.

4.6 Principle 6: a binding resolution can signal before implementation

Let $\Omega = \{I, J\}$, where I means implementation and J nonimplementation. Suppose the participant's beliefs match the announced probability: $\rho(I) = \pi \in (0, 1]$ and $\rho(J) = 1 - \pi$. A binding resolution fixes the action before this uncertainty is resolved. Under the additive specification, if implementation gives x_a and nonimplementation gives the same direct consequence x_0 for every action, then

$$U_\pi(a,q) = \pi\, u(x_a,q) + (1-\pi)u(x_0,q) + d_\pi(a). \tag{16}$$

The subscript π records dependence on the implementation probability. Diagnostic value is received when the binding choice is made, so it has no multiplying π. This states BP's timing argument (pp. 14–15) and does not require affinity of D.

Interpretation can still change with π: changing the probability can change which dispositions choose each action. Thus $d_\pi(a)$ need not be constant. If merely choosing an action has a direct moral cost, its nonimplementation consequence can also differ across actions. The common-consequence assumption in this illustration would then need to be relaxed. The guarantee decision in Section 7 has its own payoff structure, but uses the same distinction between interpreting a binding choice and observing its later outcome.

5 Restrictions that follow from the principles

5.1 A conditional sure-thing comparison

Under the additive specification, let $E \subseteq \Omega$ be an event and let $f_E h$ give $f(\omega)$ on E and $h(\omega)$ outside E. Hold q, ρ, utilities, and the two posteriors fixed. Equation (9) implies

$$(f_E h, \mu) \succeq_q (g_E h, \nu) \quad \Longleftrightarrow \quad (f_E k, \mu) \succeq_q (g_E k, \nu). \tag{17}$$

Both comparisons have utility difference

$$\sum_{\omega\in E} \rho(\omega)\{u(f(\omega),q) - u(g(\omega),q)\} + D(\mu) - D(\nu).$$

The common direct consequences outside E cancel. This follows Savage's sure-thing reasoning while keeping interpretation fixed. When $\mu \neq \nu$, the complete consequences outside E differ, so this is not the full sure-thing axiom on complete consequences. Recalculating posteriors after changing the acts would be a different comparison.

5.2 Bayesian consistency and uninformative choices

Proposition 1 (Prior preservation)**.** *Under true interpretations,*

$$\sum_a m(a)\mu_a(\theta) = F(\theta) \quad \textit{for every } \theta \in Q. \tag{18}$$

If $\sigma(a \mid q)$ is independent of q, every positive-probability action has posterior F.

Proof. For actions with $m(a) > 0$, multiply (14) by $m(a)$ and sum over actions. Unused actions contribute zero. Since $\sum_a \sigma(a \mid \theta) = 1$, the result is $F(\theta)$. If all dispositions have the same action likelihood, it cancels from Bayes' rule. □

The average posterior equals the prior under the same subjective model. This result requires no numerical diagnostic utility. If the additive specification also uses affine D, then

$$\sum_a m(a)d(a) = D\left(\sum_a m(a)\mu_a\right) = D(F). \tag{19}$$

For nonlinear D, the average diagnostic value need not equal $D(F)$. Prior preservation therefore survives in the ordinal model, while constant average diagnostic value is a restriction of BP's affine specification.

Neither identity eliminates signaling incentives: each disposition compares actions with their different posteriors. Choosing an entire rule and recalculating all posteriors is a different decision problem. For any Bayesian choice rule with the same prior, affinity gives the same subjective average diagnostic value. This is an average across possible actions and types, not the value of the action actually selected by a particular type.

The weights $m(a)$ are subjective predictions. Using observed choice frequencies in their place requires a justified link between the prior, the participant population, and the belief measures. Under face-value interpretations, the identities need not hold when weighted by actual choices.

5.3 Personal rules and the moral-placebo comparison

Under true interpretations, if all dispositions choose the same action a^*, its posterior equals the prior. This is called *pooling*. Given posteriors for unused actions, it is optimal exactly when $(f_{a^*}, F) \succeq_q (f_b, \mu_b)$ for every q and b. Under the additive specification, this becomes

$$G(b,q) - G(a^*,q) \leq D(F) - d(b) \qquad \text{for every } q \in Q,\ b \in \mathcal{A}. \tag{20}$$

The direct gain from deviating cannot exceed the diagnostic loss. A personal rule can therefore be followed even when following it supplies no new information about disposition (BP, pp. 12–13).

A more favorable prior in the diagnostic ranking weakly raises $D(F)$ under a compatible additive representation. If direct utilities and the diagnostic values of unused actions stay fixed, these inequalities become easier to satisfy. This gives a conditional version of BP's moral-placebo argument (pp. 16–17). It concerns whether a pooling equilibrium can be sustained. It does not select an equilibrium or imply that a more favorable self-image always increases helping. BP also qualifies the behavioral predictions for discrete actions (p. 14).

6 From the model to observed choices

The experiment below compares paying to guarantee relief from a task with leaving that relief to chance. Write $a = 1$ for purchasing the guarantee and $a = 0$ for accepting the lottery. Under the additive representation, the participant weakly prefers the guarantee exactly when

$$G(1,q) - G(0,q) + D(\mu_1) - D(\mu_0) \geq 0. \tag{21}$$

The first difference is the direct advantage of the guarantee. It includes its price, its effect on the probability of avoiding the task, and any direct value of controlling the outcome. The second difference is its diagnostic advantage. Both actions can convey information: declining to pay is not assumed to leave self-beliefs unchanged.

One prediction can be stated without assigning numerical values to self-image. Hold the direct acts fixed, and compare the original interpretations (μ_1, μ_0) with new interpretations (ν_1, ν_0).

Proposition 2 (A better diagnostic comparison)**.** *If $\nu_1 \succeq^D \mu_1$ and $\mu_0 \succeq^D \nu_0$, a weak preference for the guarantee is preserved. An existing strict preference is also preserved.*

Proof. The common diagnostic ranking and the original preference imply

$$(f_1, \nu_1) \succeq_q (f_1, \mu_1) \succeq_q (f_0, \mu_0) \succeq_q (f_0, \nu_0).$$

Transitivity gives the result. If the middle comparison is strict, the resulting comparison is strict. □

This is a conditional theoretical prediction. It does not directly compare helping another person with avoiding one's own task, because those decisions have different direct consequences. The experiment therefore retains both the direct benefit and the diagnostic benefit in its quantitative specification.

Timing also matters. A person can interpret the decision before knowing the lottery outcome. The resulting diagnostic value is not automatically multiplied by the probability that relief occurs. This does not mean that interpretation is constant across probabilities or prices: a costly decision, or one that produces only a small improvement in another person's circumstances, may reveal something different. Anticipated beliefs must therefore refer to the exact choice being offered. A direct reward from the act of helping, independently of any belief about disposition, is another possible motive. Unless it is restricted or measured separately, it can be confused with diagnostic utility.

7 Experimental design

7.1 The decision and the two conditions

The experiment asks whether people value a helping decision partly because it makes them see themselves as considerate. Participants can pay to replace an uncertain outcome with a guaranteed one. Following Feldhaus et al. [3], a separate monetary lottery provides a benchmark for how much that guarantee is worth in money alone. Belief questions then measure what the decision reveals about the participant. The design is a proposal; no experimental results are reported here.

The task is two minutes of repetitive proofreading: identifying typographical errors in short texts. Participants see a brief practice example before deciding. The task lasts for the stated time rather than ending when a target score is reached. Whoever is relieved of the task has those two minutes free, and their fixed participation payment is unchanged. Whether participants are willing to pay for this relief is an empirical question.

Participants are randomly assigned to one of two conditions. In the *helping condition*, D, the decision concerns an anonymous recipient's task; the decision-maker's own workload is fixed. In the *control condition*, N, it concerns the decision-maker's own task; no other person's workload changes. Recipients are recruited separately and never learn who made the decision. Decisions and belief reports are private. The control offers a real benefit—avoiding one's own effort—through the same procedure. It is not assumed to have zero diagnostic value; the belief questions examine that possibility.

The conditions have identical monetary payments, task duration and probabilities. However, relieving someone else's task and relieving one's own task need not have the same direct value. Consequently, a difference in willingness to pay between D and N is not, by itself, a measure of diagnostic utility.

7.2 Choices, payments and timing

Each participant faces one price C and one probability π, both assigned at random. The probability is either 25% or 75% and refers to the chance of receiving EUR 8 *and* avoiding the relevant task if the participant does not buy the guarantee. The price ranges from EUR 0 to EUR 6 in EUR 0.50 increments. The screen presents the two alternatives as follows.

Decision	**What happens**
Buy the guarantee	The participant receives EUR $(8 - C)$ for certain. The relevant person does not do the task.
Accept the lottery	With probability π, the participant receives EUR 8 and the relevant person does not do the task. Otherwise, the participant receives EUR 4 and that person completes the two-minute task.

In D, the relevant person is the recipient; in N, it is the participant. For example, at $C = 2$ and $\pi = 25\%$, buying the guarantee gives the participant EUR 6 and removes the task for certain. Accepting the lottery gives a one-in-four chance of EUR 8 with no task and a three-in-four chance of EUR 4 with the task. The EUR 8 and EUR 4 are lottery payments to the decision-maker, not wages for proofreading. All participants also receive a fixed participation payment, and the recipient's payment does not depend on the decision. The two lottery outcomes are displayed together so that the monetary and workload consequences cannot be mistaken for independent draws.

Before choosing, participants answer short comprehension questions about who would do the task, what the stated probability concerns, and what payment each outcome produces. They then report their beliefs as described below, make the binding choice, and report their self-belief once more. There are no repeated rounds of the main decision. Prices vary across participants, so the experiment can estimate how the frequency of buying the guarantee changes with its price without treating several decisions by the same person as independent signals.

After a short neutral filler task, participants complete the monetary benchmark. It uses the same EUR 8/EUR 4 lottery and their assigned probability π, but it has no effect on anyone's workload. Participants choose between this lottery and sure amounts from EUR 4 to EUR 8 in EUR 0.25 increments. One row is selected at random and paid in addition to the main decision. If choices switch once, the switch brackets the certainty equivalent, CE_π: the sure payment judged as good as the monetary lottery. Multiple switches are retained as inconsistent responses rather than forced into a single value. Interpreting this task as a monetary benchmark assumes that its presentation creates no additional diagnostic or procedural premium, and that monetary preferences are sufficiently stable across the two tasks.

All choices and belief reports are collected before the main lottery is drawn, payments are revealed or proofreading begins. The monetary benchmark has its own payment account and random draw. Random selection of its paid row is treated as incentive compatible under the maintained expected-utility assumptions. It does not introduce another chance of implementing the main choice: that choice is always binding.

7.3 What participants believe the choice reveals

The belief questions distinguish concern for others from a person's assessment of that concern. Let H denote stronger concern for sparing others unnecessary effort and L weaker concern. These are two possible underlying preferences, not labels assigned according to the current decision. A person of either type may pay, because money, the recipient's relief and self-image can all

matter. The two-type description is a simplifying assumption; the experiment does not observe a participant's true type.

Before choosing, participants read both descriptions and answer: *"How likely is it that the description 'stronger concern for sparing others unnecessary effort' fits you better than the description 'weaker concern'?"* They respond from 0 to 100. Dividing the answer by 100 gives the reported prior, $F(H)$. The same question is then asked under each of two separate possibilities: *"Suppose you buy the guarantee at this price"* and *"Suppose you accept the lottery instead."* These answers measure the anticipated beliefs $F(H \mid 1)$ and $F(H \mid 0)$. Both are needed because the model compares the interpretations of the two actions. After the binding decision, the original question is repeated, still before any lottery outcome. This provides a check on whether the belief actually reported after choosing agrees with the anticipated belief for that action.

The questionnaire also asks which description the participant would prefer to apply to themselves. All belief questions receive a fixed payment rather than a reward for accuracy: there is no independently observed true disposition against which to score them. Reports preferring L are recorded as disagreements with the maintained ranking of dispositions, rather than silently treated as preferences for H.

After the main choice, participants complete a short exercise about someone they know, without naming that person. They assess the probability that the stronger-concern description fits that person, and then reassess it under each hypothetical decision at the same price and probability. The two scenarios are separate alternatives, not successive decisions; their order is randomized. This exercise makes the interpretation of an action more concrete. It can show, for example, whether paying is seen as evidence of concern and whether declining to pay is seen as evidence against it. It measures beliefs about another person and does not automatically replace a report about oneself. In particular, the friend's initial assessment may differ from the participant's self-assessment.

Finally, participants predict how many out of 100 people of each type would buy the guarantee at the same price and probability. These two reports allow a comparison between the directly reported posterior and the posterior calculated by Bayes' rule, as described in Appendix A.1. Using judgments about others to interpret one's own action requires an additional assumption that the action has the same meaning in both cases.

7.4 How diagnostic utility enters the comparison

The empirical specification uses BP's additive and affine case from Section 4:

$$\begin{aligned} U(a,q) &= G(a,q) + \sum_{\theta\in\{H,L\}} F(\theta \mid a)V(\theta), \\ V(L) &= 0, \qquad V(H) = \lambda. \end{aligned} \tag{22}$$

Here $\lambda > 0$ is the utility difference between certainty of being H and certainty of being L, under the maintained strict preference for H. A change in a belief is information about self-image; it is not yet a measure of its utility. The coefficient λ tells us how much that information matters for choice. The common ordinal ranking does not itself determine a common monetary λ.

The first empirical object is willingness to pay for the guarantee. If a person's choices could be summarized by an indifference price $\bar{C}^c_\pi$, the Feldhaus-style monetary correction would be

$$\begin{aligned} \widehat{V}^c_{\pi,\text{raw}} &= \bar{C}^c_\pi - (8 - CE_\pi), \\ \widehat{V}_\pi &= \widehat{V}^D_{\pi,\text{raw}} - \widehat{V}^N_{\pi,\text{raw}}. \end{aligned} \tag{23}$$

The amount $8 - CE_\pi$ is the willingness to pay for the monetary improvement alone. For example, if $CE_\pi = 5$ and the guarantee is worth EUR 4, the corrected premium is EUR 1. This premium is

a monetary equivalent of everything the guarantee adds beyond the monetary lottery, including task relief and any value of control or self-image. It is distinct from the type-value function $V(\theta)$ and is not automatically diagnostic utility.

To separate these motives in the simplest numerical specification, assume locally linear monetary utility, additive task values, and beliefs that agree with the announced lottery probabilities. Let b_q^D be the direct value of relieving the recipient's task and b_q^N that of relieving one's own task, expressed in euros. Assume these values remain stable across the two probabilities. Also assume that the ordinary value of guaranteeing implementation is common to D and N at the relevant comparisons, so it cancels, and that the monetary diagnostic coefficient λ is stable. Writing

$$\Delta F_\pi^c = F^c(H \mid 1) - F^c(H \mid 0),$$

the difference in corrected premiums satisfies

$$\widehat{V}_\pi = (1-\pi)(b_q^D - b_q^N) + \lambda(\Delta F_\pi^D - \Delta F_\pi^N). \tag{24}$$

Each posterior in this equation is evaluated at its condition's indifference price. The direct-benefit term cannot be dropped merely because monetary payments are matched. Nor is the control's belief gap assumed to be zero.

The two probabilities help because buying the guarantee increases the chance of relief by 75 percentage points when $\pi = 25\%$, but by only 25 percentage points when $\pi = 75\%$. A stable direct value of relief therefore contributes three times as much in the first case. Diagnostic value need not follow that proportion: it depends on the measured difference in interpretations. Separation is possible only if the measured belief difference does not change in that same proportion and the stated preference restrictions hold. Appendix A.2 gives the algebra.

These indifference-price equations explain the model; one observed decision per participant does not reveal an individual willingness to pay. Estimation must therefore use the observed choices at their assigned prices and specify how preferences vary across participants. Random assignment makes the groups comparable on average, but does not make participants identical or reveal their types. Appendix A.2 states the corresponding choice comparison. When several valuations fit the evidence, the result is a range rather than a unique estimate of λ.

The quantitative interpretation is conditional. An extra direct reward from helping, or a different value of control across conditions, could otherwise be mistaken for diagnostic utility. With nonlinear monetary utility, a certainty equivalent alone does not supply the utility differences needed for the decomposition; a separate calibration would be required. Belief reports can also contain error, and asking about self-image before choosing may make it more salient. Questions and timing are therefore identical across conditions, and conclusions concern this elicitation procedure.

The experiment can consequently establish how the opportunity to help changes choices and reported self-beliefs relative to avoiding one's own effort. It can assess whether those observations are compatible with a diagnostic-utility explanation under explicit restrictions. It does not identify diagnostic utility from willingness to pay alone, and reported likelihoods do not establish BP's stronger true-interpretations condition without evidence linking actual dispositions to behavior.

8 Conclusion

Diagnostic utility adds a value for what an action reveals about the chooser to the value of its direct consequences. The framework separates the underlying disposition, beliefs about that disposition, and preferences over those beliefs. It states which conclusions require only an ordering of self-beliefs and which require an additive or affine numerical representation. Bayesian updating supplies a consistency condition for interpretations; it does not derive preferences, identify the prior, or by itself select a unique pattern of behavior.

The proposed experiment connects these objects to a concrete decision: paying to spare an anonymous person a task, compared with paying to avoid that task oneself. A monetary lottery provides the freedom-of-choice benchmark, while explicit prior and posterior questions measure the interpretation of both alternatives. Varying the probability of relief changes its direct benefit in a known proportion. Separating diagnostic value requires the measured belief contrast to vary differently, together with stable preferences and a justified account of heterogeneity. When these conditions are not supported, the experiment still measures choices and reported self-beliefs, but does not supply a separate numerical estimate of diagnostic utility. This distinction keeps the proposed test consistent with the scope of the theoretical model.

A Belief consistency and conditional identification

A.1 The Bayesian comparison

The two likelihood questions ask how many out of 100 people of type H, and of type L, would buy the guarantee in the stated situation. Dividing by 100 gives $\Pr(1 \mid H)$ and $\Pr(1 \mid L)$. The likelihoods of accepting the lottery are their complements. With $F(L) = 1 - F(H)$, the implied posterior is

$$F(H \mid a) = \frac{F(H)\Pr(a \mid H)}{F(H)\Pr(a \mid H) + F(L)\Pr(a \mid L)}, \qquad a \in \{0, 1\}, \tag{25}$$

whenever the denominator is positive. No posterior is calculated from a zero denominator. Equal action likelihoods for the two types imply no change from the prior.

The formula can be compared with the directly reported anticipated posterior. Applying likelihoods about others to one's own prior requires a shared interpretation of the action. Disagreement may reflect a failure of that assumption, departures from Bayesian reasoning, or reporting error. Moreover, the likelihoods are collected after the choice and may be affected by it. This is a check of the consistency of reports, not independent verification of actual type-dependent behavior. The social-circle exercise is interpreted in the same way: it concerns the meaning attached to a known person's choice, rather than identifying the participant's own disposition.

A.2 What the two probabilities can separate

First consider a fixed participant with well-defined indifference prices in both conditions and both probability settings. Under the restrictions in Section 7, the direct-relief difference in equation (24) is multiplied by .75 in one setting and .25 in the other. It therefore cancels in

$$\begin{aligned}\widehat{V}_{.25} - 3\widehat{V}_{.75} = \lambda\Big[&(\Delta F^{D}_{.25} - \Delta F^{N}_{.25}) \\ &- 3(\Delta F^{D}_{.75} - \Delta F^{N}_{.75})\Big].\end{aligned} \tag{26}$$

If the bracket is nonzero, this equality recovers λ by division. A negative implied value would contradict the maintained preference for H, subject to measurement error and the other restrictions. If the bracket is zero, the comparison does not recover λ: a zero left-hand side is compatible with multiple values, while a nonzero left-hand side contradicts the maintained specification. All belief gaps here refer to the appropriate indifference prices; they cannot be replaced by beliefs elicited at unrelated prices.

The actual design observes one posted-price decision per participant. Under linear monetary utility its utility difference is

$$\begin{aligned}U^c(1, q) - U^c(0, q) = 4(1 - \pi) - C + (1 - \pi)b^c_q + k_\pi \\ + \lambda\big[F^c(H \mid 1) - F^c(H \mid 0)\big].\end{aligned} \tag{27}$$

Here k_π is the ordinary value of guaranteeing implementation, assumed common across D and N for this comparison. The first term is the monetary gain from replacing the EUR 8/EUR 4 lottery; under linear utility, $CE_\pi = 4 + 4\pi$. Buying requires the displayed difference to be nonnegative; accepting the lottery requires it to be nonpositive. The anticipated beliefs now refer to the participant's actual posted price. If the data require stochastic choice, its source and distribution must be specified rather than read from these deterministic inequalities.

The cutoff calculation establishes conditional identification in the fixed-parameter model. It does not show that the proposed data identify all parameters when types and valuations differ across people. A population analysis needs explicit restrictions on that heterogeneity and on belief-reporting error, or must report sets of compatible parameters. It cannot substitute a group's cutoff for an individual's unobserved cutoff or equate a subjective self-prior with the population distribution of actual types. Estimating that richer model is separate from establishing the identification logic of the experiment.

References

[1] Bodner, R., and D. Prelec (2001). *Self-signaling and diagnostic utility in everyday decision making*. May 2001 manuscript; page references in the text refer to this version. Published in 2003 in I. Brocas and J. D. Carrillo (eds.), *The Psychology of Economic Decisions*, Vol. 1, pp. 105–124. Oxford University Press. doi:10.1093/oso/9780199251063.003.0006.

[2] Savage, L. J. (1954). *The Foundations of Statistics*. Wiley. Second revised edition, Dover, 1972.

[3] Feldhaus, C., J. Lingens, A. Löschel, and G. Zunker (2026). Measuring the intrinsic value of choice. *European Economic Review*, 184, 105254. doi:10.1016/j.euroecorev.2025.105254.

[4] Bénabou, R., and J. Tirole (2004). Willpower and personal rules. *Journal of Political Economy*, 112(4), 848–886. doi:10.1086/421167.

[5] Bénabou, R., and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5), 1652–1678. doi:10.1257/aer.96.5.1652.

[6] Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401), 464–477. doi:10.2307/2234133.

[7] Gächter, S., L. Molleman, and D. Nosenzo (2025). Why people follow rules. *Nature Human Behaviour*, 9, 1342–1354. doi:10.1038/s41562-025-02196-4.

[8] Gneezy, U., A. Kajackaite, and J. Sobel (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2), 419–453. doi:10.1257/aer.20161553.