# Hierarchical Aggregation of Information and Decision-Making *

Tom Cunningham†

August 17, 2015

**Abstract**

It is commonly argued that the brain aggregates information in a hierarchical fashion. In this paper I point out that hierarchical aggregation of information will give rise to certain predictable imperfections in inference, consistent with well-known features of perceptual illusions and decision biases. I consider a setup with two modules: both infer some unobserved value, each has private information, but the second module additionally observes the first module's posterior. As a whole this system will fail to aggregate information efficiently, in particular it predicts two particular features of decision-making: (1) the influence of irrelevant associations (framing effects), and (2) the avoidance of dominated options. This combination of properties distinguishes the model from either random preferences or inattention.

*Previously circulated as "Biases and Implicit Knowledge". Among many others I thank for their comments Roland Benabou, David Cesarini, Erik Eyster, Scott Hirst, Dorian Jullien, David Laibson, Vanessa Manhire, Arash Nekoei, José Luis Montiel Olea, Robert Östling, Alex Peysakhovich, Jon de Quidt, Ariel Rubinstein, Benjamin Schoefer, Andrei Shleifer, Rani Spiegler, Dmitry Taubinsky, Matan Tsur, Michael Woodford, and seminar participants at Harvard, Tel Aviv, Princeton, HHS, the IIES, Oxford, USC Marshall, WZB Berlin, and the Paris School of Economics. I acknowledge financial support from ERC grant 269143.

†IIES, Stockholm University, tom.cunningham@iies.su.se.

This paper studies the problem of hierarchical aggregation of information. Suppose that two agents are both trying to infer some value $v$, that each agent receives a private signal, and that the second agent additionally observes the posterior belief of the first agent regarding $v$. We can then ask under what conditions this *hierarchical* aggregation is efficient, i.e. under what conditions is the second agent's belief equal to the belief they would have formed if they had access to both private signals? I additionally study how the answer changes when there are two or more values to be judged simultaneously ($v$ and $v'$, etc.) meaning that the second agent will be able to infer more of the first agent's private signal from observing two of his posteriors.

The second contribution of this paper is to suggest that hierarchical aggregation of information could explain some puzzling features of human decision-making. Hierarchical processing is already used, informally, to explain a variety of optical illusions and other perceptual biases.[1] Such explanations are based on the existence of dedicated perceptual systems which have limited inputs, and therefore form inferences which are biased relative to the full-information case. I show how this case can be formalized, and additionally use hierarchical processing to explain value-illusions just as it is used to explain visual illusions.

Put in ordinary language the basic analysis is that our decisions are based on intuitions, and those intuitions are informative but coarse. When we evaluate a bottle of wine, a house, or a job candidate, we receive an overall feeling, with imperfect insight into how that feeling is generated. The feeling is informative, meaning it is correlated with the underlying value, but it is also coarse in the sense that it is based only on a limited set of salient inputs. This analysis can explain why our decisions are inconsistent between contexts: because the intuitions are coarse. But also why our decisions are consistent within contexts: because we are able to reflect on the intuitions, and impose order on them. More specifically the decision-making process will exhibit two features:

1. *Influence of irrelevant associations.* Decisions will be influenced by features that the decision-maker knows to be irrelevant, i.e. framing effects will occur. Moreover, the influence of a feature will depend on that feature's

[1] Fodor (1983) and Pylyshyn (1984) gave influential accounts of illusions as illustrating the encapsulation of mental systems; Adelson (2000) gives many elegant examples of illusions as optimal inferences.

association with value in the experience of the decision-maker. I discuss a broad range of evidence for such effects in a later section.

2. *Local consistency.* Decisions will obey normative restrictions individually but not collectively. For example, the model predicts that people will never directly choose a dominated option, even though their choices could indirectly reveal preference for a dominated option. This fits a stylized fact of the experimental decision-making literature: while decisions are often noisy and biased, it is comparatively rare for people to choose a dominated option.

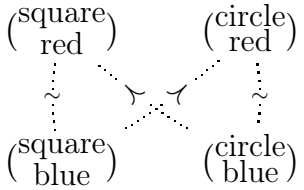
These two characteristics sharply distinguish the model from two other popular accounts of inconsistent decision-making. In *inattention* models decision-makers imperfectly perceive the choice set, and so exhibit biases relative to the full-information case. However such models also predict that, because decision-makers perceive options imperfectly, they will frequently choose dominated alternatives. In *random preference* models decision-makers imperfectly perceive their preferences. These decision-makers can make inconsistent choices, while never violating dominance. However such models do not predict the existence of systematic biases such as the influence of an attribute known to be irrelevant. The hierarchical model in this paper can thus be thought of as a combination of inattention and random preferences.

As a concrete example, consider evaluating one of two offers: either a children's toy that costs $10, or the same toy at $7 with $3 shipping. Normatively the two offers are identical, yet a variety of studies find that people treat purchases differently depending on the partition.[2] This behaviour would be consistent with hierarchical processing of information if each of the two partitions evoked different associations, for example if free shipping evoked a positive association. The positive feeling generated will be attributed, in part, to the value of the toy itself. The same behaviour could be explained by inattention, but hierarchical processing makes an additional testable prediction: that the decision-maker would be indifferent between the two offers when evaluated side by side.

A characteristic pattern in choice predicted by this model is a "figure 8" intransitivity. Suppose we observe the following binary choices over two attributes,

[2]Chetty et al. (2009), Brown et al. (2010).

shape and colour:




These preferences describe a decision-maker who indifferent between red and blue, whenever the shape is the same, but who strictly prefers blue over red whenever the shapes differ. These choices are intransitive, so cannot be generated by maximizing a stable utility function. Analyzed using the model of hierarchical aggregation of information these choices imply two facts: (1) the second system believes colour to be irrelevant; (2) the first system has a positive association with blue, unknown to the second system.

There are a number of novel economic applications. The existence of implicit knowledge predicts the existence of uninformative advertising: firms will wish to evoke positive associations when presenting their product, because the decision-maker relies on an overall feeling about the product, and they cannot fully discount irrelevant attributes which contribute to that feeling. The model also predicts that true preferences can be inferred from inconsistent choices, and predicts that the quality of choices can be improved by providing decision-makers with additional alternatives for comparison.

In the following sections I, first, state the general model and its implications; second, state the parametric model and its implications; third, discuss existing evidence and new predictions; and finally, discuss related literature and economic applications. An appendix contains supplementary discussion and proofs.

# 1 Model

The formal model consists of two systems, which operate sequentially, called "System 1" and "System 2", as in Figure **??** (terminology borrowed from Stanovich and West (2000)). Each system wishes to estimate an unobserved value $v$ from the observed details $x$. System 1's evaluation is imperfect relative to System 2, because it does not take into account all aspects of the case, i.e. it receives only a coarse signal $s(x)$. Nevertheless System 2 is rationally influenced by System 1's

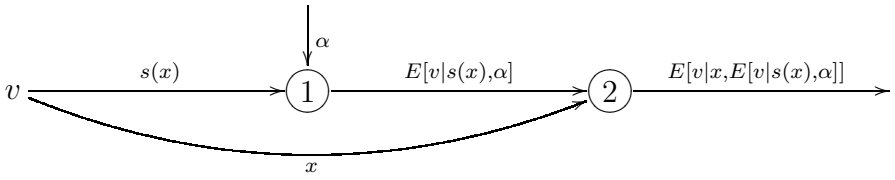



**Figure 1:** A hierarchical model: System 1 receives $s(x)$ and $\alpha$; System 2 receives $x$ and System 1's expectation of $v$. In most applications $\alpha$ will be assumed to be a constant: this implies that if System 2 can observe System 1's expectations for multiple cases $(x^1, x^2, \ldots)$ it will be able to learn more about $\alpha$, and so bias will fall.

judgment, because System 1 has access to information not available to System 2, associations summarized by $\alpha$. Formally, the model can be interpreted as a contribution to the theory of social learning, which studies conditions under which private information is efficiently aggregated by a sequence of agents, each of whom observes earlier agents' actions (Chamley (2003)).

In this section I show (1) sufficient conditions for the efficient aggregation of information; (2) how the solution changes with multiple evaluations; (3) the local consistency of evaluations and decisions; and (4) the influence of irrelevant attributes.

Assume a probability space $(\Omega, E, P)$, and three random variables $v \in \mathbb{R}$, $x \in X$, $\alpha \in A$, defined by the measurable functions $F_v : \Omega \to \mathbb{R}$, $F_x : \Omega \to X$, and $F_\alpha : \Omega \to A$. I define a joint distribution measure $f(v, x, \alpha) \equiv P(\{\omega | F_v(\omega) = v, F_x(\omega) = x, F_\alpha(\omega) = \alpha\})$, from which conditional distributions are derived. Finally I define a function $s : X \to S$, which represents the coarse signal $s(x)$ which System 1 receives about the case $x$.

We can then define the following expectations, which represent respectively the expectations about $v$ formed by System 1, by System 2, and by a hypothetical agent who is able to pool both information sets:

$$\begin{aligned} E_1 &= E[v|s(x), \alpha] \\ E_2 &= E[v|x, E[v|s(x), \alpha]] \\ E_P &= E[v|x, \alpha]. \end{aligned}$$

In general $v$ will be interpreted as the consumption-value of a particular prospect, something that it is unknown and must be estimated from the observable features; $x$ will be interpreted as information about the current case, with $s(x)$ representing superficial features, and the information not in $s$ representing more abstract or "high-level" features; finally $\alpha$ will be interpreted as static knowledge held by System 1. A simple interpretation of $\alpha$ is as the history of the decision-maker's previous experiences in similar situations, i.e. the historical association between each $s \in S$ and $v$. Interpretation of $\alpha$ is discussed further in the appendix, and a learning foundation is provided. In the example of seeing apartments, $x$ represents all the details about the apartment; $s(x)$ includes only superficial features, such as the weather on the day of the visit, but does not include high-level information, like the fact that the weather is irrelevant when buying an apartment. Finally $\alpha$ represents System 1's implicit knowledge, such as the agent's learnt association between weather and satisfaction.

The literature on decision-making does not have an agreed-upon formal definition of the word "bias", however an informal definition could be that it is an avoidable misjudgment, i.e. a failure to use all the information that was available. In this model there exist quantities which can naturally be described as representing biases, in this sense: for each agent, the difference between their expectation and the pooled-information expectation:

$$\begin{aligned} \text{System 1's bias} &= E_1 - E_P \\ \text{System 2's bias} &= E_2 - E_P \end{aligned}$$

It should be clear that both $E_1$ and $E_2$ will have a zero average bias, i.e. $E[E_1 - E_P] = E[E_2 - E_P] = 0$, however System 2's bias will always be smaller in expectation:

*Remark* 1. System 2 has a smaller average bias (by mean squared error)[3]

$$E\left[(E_2 - E_P)^2\right] \leq E\left[(E_1 - E_P)^2\right].$$

The paper will be mainly interested in the properties of $E_2$, and when I say

[3]This follows from the fact that the variance of an expectation's error will be smaller if it conditions on more information: For any random variables $v$, $p$, $q$, $Var[v - E[v|p,q]] \leq Var[v - E[v|p]]$.

that judgment is unbiased I will mean that for all $x \in X, \alpha \in A$,

$$\begin{aligned} E_2 &= E_P \\ E[v|x, E[v|s(x), \alpha]] &= E[v|x, \alpha]. \end{aligned}$$

The model could be generalized to include a rule which determined when to activate System 2: for example, when stakes are low, then decisions will be made automatically using $E_1$; when stakes are high, then System 2 is also activated and $E_2$ is calculated.[4]

In fact the brain surely is more complex than just two sequential systems. What is important for this model is that even with maximum mental effort not all information is efficiently aggregated, i.e. there are some things that we know, and affect our decisions, but to which we do not have conscious access.

In the remainder of this section I show that, with implicit knowledge, (1) judgments will be unbiased under certain conditions on the joint distribution of $x$ and $\alpha$; (2) bias will be smaller when considering a larger set of cases (extending the model to a set of cases $\mathbf{x} = (x^1, \ldots, x^m)$ judged simultaneously); (3) any set of judgments made jointly will be internally consistent (i.e., can be justified by some beliefs over $\alpha$); and (4) judgments will be influenced by irrelevant associations, under an assumption about the symmetry of the unobserved information $\alpha$.

## 1.1 Conditions for Existence of Bias

In this section the shared information ($s(x)$) is not important, instead we concentrate on the information that is private to each system. The key result is that bias will occur only when there is an *interaction* between each system's information ($\alpha$ and $x$), meaning, roughly that System 1's information cannot be interpreted without knowing System 2's information.

A simple sufficient condition for unbiasedness is that $E[v|s(x), \alpha]$ is a one-to-one function of $\alpha$; then System 2 can immediately infer $\alpha$. This condition is not necessary: in many cases System 2 can extract all the information it needs from $E_1$ without knowing $\alpha$. The relationship between $E_1$, $E_2$, and $E_P$ can be illustrated in the following table (here the shared information is assumed to be

[4]See Brocas and Carrillo (2013) for a related setup.

constant, i.e. $s(x) = s(x') = s$):

| | $\alpha$ | $\alpha'$ | $\alpha''$ |
|---|---|---|---|
| $x$ | $E[v\|x,\alpha]$ | $E[v\|x,\alpha']$ | $E[v\|x,\alpha'']$ |
| $x'$ | $E[v\|x',\alpha]$ | $E[v\|x',\alpha']$ | $E[v\|x',\alpha'']$ |
| | $E[v\|s,\alpha]$ | $E[v\|s,\alpha']$ | $E[v\|s,\alpha'']$ |

Each column represents a different realization of $\alpha$, and each row represents a realization of $x$. The six interior cells correspond to the pooled expectation, $E_P$, under different realizations of $\alpha$ and $x$. The elements of the last row correspond to $E_1$, i.e. they are average expectations conditioning only on $\alpha$. Finally, the two outlined cells correspond to a realization of $E_2$, i.e. a set of cells in a row grouped according to whether their columns share the same $E_1$: here the border is drawn under the assumption that $E[v|s,\alpha] = E[v|s,\alpha'] \neq E[v|s,\alpha'']$.

A bias occurs when $E_2 \neq E_P$, thus in the table it will occur when the rectangle representing $E_2$ includes cells with different values. A necessary and sufficient condition for unbiasedness is therefore that any two columns which have the same average expectation ($E_1$) must also have the same expectation for every $x$.

*Remark* 2. Judgment will be unbiased if and only if, for all $x \in X$, $\alpha, \alpha' \in A$,

$$E[v|x,\alpha] \neq E[v|x,\alpha'] \implies E[v|s(x),\alpha] \neq E[v|s(x),\alpha'].$$

For example aggregation would fail if $\alpha$ and $x$ were real numbers, distributed independently, with $E_P = \alpha x$, $E[x] = 0$, and $S$ is a singleton (i.e., $s(x)$ is uninformative). In this case System 1 does not know whether any given realization of $\alpha$ represents better or worse news about $v$: i.e., no matter the realization of $\alpha$, $E_1 = 0$, and for every $x$, $E_2 = 0$, implying that judgment will be biased ($E_2 \neq E_P$) whenever $\alpha \neq 0$ and $x \neq 0$.[5]

A stronger condition for unbiasedness exists when $\alpha$ and $x$ are independent. Independence of $\alpha$ and $x$ may be reasonable when $\alpha$ is interpreted as long-run knowledge (knowledge about how to interpret $s(x)$) and $x$ represents idiosyncratic features of the current case. Under independence judgments will be almost surely unbiased if $\alpha$ is monotonic, in the sense that a change from $\alpha$ to $\alpha'$ is either always

[5]A fuller version of this example is given in the appendix.

good news or always bad news (i.e., it either always weakly increases or weakly decrease the expected $v$ for every $x$).

**Proposition 1.** *Judgment will be almost surely unbiased if $\alpha$ and $x$ are independent and, for every $s \in S$, there exists some total order $\succeq_s$ on $A$, such that for all $x \in X$, $E[v|x, \alpha]$ is weakly monotonic in $\alpha$ when ordered by $\succeq_s$.*

In terms of the table above, $\alpha$ is monotonic if the columns can be rearranged in such a way that the elements in every row are weakly increasing.

It follows from proposition 1 that no bias will occur when $E_P$ is a separable function of $\alpha$ and $x$, i.e. there must be some *interaction* between the two pieces of information for bias to occur.

**Corollary 1.** *Judgment will be unbiased if $\alpha$ and $x$ are independent, and there exist functions $g : S \times A \to \mathbb{R}$, $h : X \to \mathbb{R}$, and $i : \mathbb{R} \to \mathbb{R}$, such that*

$$E_P = E[v|x, \alpha] = i(g(s(x), \alpha) + h(x)),$$

*and $i$ is strictly monotonic.*

*Proof.* In this case for any $s(x)$ there exists an ordering of $A$ such that $E_P$ is monotonic in $\alpha$ for any $x$ (i.e., the ordering according to $g(s(x), \alpha)$). Judgment will therefore be unbiased, by the previous proposition. □

Arieli and Mueller-Frank (2013) show that there will be no bias if the signals $\alpha$ and $x$ are conditionally independent (given $v$), and if System 2 can infer from $E_1$ the entire posterior of System 1, not just their expectation (i.e. if they can infer $f(v|\alpha)$, not just $E[v|\alpha]$).[6] Arieli and Mueller-Frank also show that $E_1$ will almost always reveal the entire posterior, in a probabilistically generic sense. The latter fact will hold in the Gaussian examples below: System 2 always will be able to infer System 1's entire posterior distribution over $v$. However in most examples of interest to this paper $\alpha$ and $x$ will not be conditionally independent, and for this reason Arieli and Mueller-Frank's theorem will not apply, and a bias will remain.

[6] In their setup the common information is irrelevant, so we can treat $S$ as a singleton.

## 1.2 Multiple Evaluations

This model makes distinctive predictions about the effect of multiple cases being evaluated, because System 2 will be able to extract information about $\alpha$ from the evaluations generated by System 1 for each case.

To represent joint evaluations I consider vectors of $m \in \mathbb{N}^+$ elements, $\mathbf{v} = \mathbb{R}^m$, $\mathbf{x} = X^m$, $\boldsymbol{\alpha} = A^m$, with the joint distribution $f_m(\mathbf{v}, \mathbf{x}, \boldsymbol{\alpha})$, and $\mathbf{s}(\mathbf{x})^i = s(x^i)$. I will refer to the vector $\mathbf{x}$ as a *situation*, and an element $x^i$ as a *case*. I assume that System 1 forms their expectations about each case as before, and that System 2 conditions each of their judgments on the entire set of expectations received from System 1, i.e. the expectations are themselves vectors, defined as:

$$\begin{aligned} \mathbf{E}_1 &= E[\mathbf{v}|\mathbf{s}(\mathbf{x}), \boldsymbol{\alpha}] \\ \mathbf{E}_2 &= E[\mathbf{v}|\mathbf{x}, \mathbf{E}_1] \\ \mathbf{E}_P &= E[\mathbf{v}|\mathbf{x}, \boldsymbol{\alpha}]. \end{aligned}$$

Applying these predictions to data requires making assumptions about when different cases belong to the same situation. When studying decision-making, a natural assumption is that the elements of the choice set together constitute a situation: i.e., when evaluating each element in the choice set, System 2 has access to System 1's evaluations for all the other members.

As written, this setup allows many channels of inference; in order to concentrate just on the channels of interest I assume that (1) all elements of $\boldsymbol{\alpha}$ are identical (as discussed above) and so refer to it as $\alpha$; (2) each case $x^i$ is distributed independently of $\alpha$; and (3) all observable information about each case is idiosyncratic, i.e. $x^i$ is informative only about $v^i$, not about $v^j$ for $j \neq i$.[7]

These three points are incorporated into the assumption that:

$$f(\mathbf{v}, \mathbf{x}, \alpha) = \left( \prod_{i=1}^{m} f(v^i|x^i, \alpha) f(x^i) \right) f(\alpha). \tag{A1}$$

Within this framework neither System 1's expectation nor the pooled-information

[7] If $x^i$ was informative about $\alpha$ or $v^j$ then we would expect joint and separate judgment to differ even without any signal from System 1.

expectation will depend on the other cases being considered,[8] but when System 2 observes a vector $\mathbf{E}_1$ it will learn about $\alpha$ from the entire set of cases, therefore any expansion in the total set of cases being evaluated will decrease the bias for any given case:

*Remark* 3. For any $m < n$, $\mathbf{x} \in X^m$, $\mathbf{x}' \in X^n$, with $x^i = x'^i$ for $i \in \{1, \ldots, m\}$, then for any $j \in \{1, \ldots, m\}$

$$Var[E_2^j - E_P^j] \geq Var[E_2'^j - E_P'^j],$$

where $\mathbf{E}_2 = E[\mathbf{v}|\mathbf{x}, E[\mathbf{v}|\mathbf{s}(\mathbf{x}), \alpha]]$ and $\mathbf{E}_2' = E[\mathbf{v}'|\mathbf{x}', E[\mathbf{v}'|\mathbf{s}(\mathbf{x}'), \alpha]]$.

This follows from the same argument as Remark 1: when conditioning on a larger information set, the variance will fall.

## 1.3 Internal Consistency

We now note that this setup implies that evaluations elicited separately can be inconsistent, but when elicited jointly they must be consistent, i.e. they can be rationalized by some beliefs $g(\alpha)$ over $A$.

*Remark* 4 (Internal Consistency). $\forall \alpha \in A$, $\forall \mathbf{x} \subseteq X^m$, $\exists g \in \mathbb{R}_+^A, \int_{\bar{\alpha} \in A} g(\bar{\alpha}) d\bar{\alpha} = 1$, such that

$$\mathbf{E}_2^\alpha = \int_{\bar{\alpha} \in A} E[\mathbf{v}|\mathbf{x}, \bar{\alpha}] g(\bar{\alpha}) d\bar{\alpha},$$

where $\mathbf{E}_2^\alpha = E[\mathbf{v}|\mathbf{x}, E[v|\mathbf{s}(\mathbf{x}), \alpha]]$.

This trivially follows in our setup, with $g(\bar{\alpha}) = f(\bar{\alpha}|\mathbf{x}, E[v|\mathbf{s}(\mathbf{x}), \alpha])$.

For example consider a pair of cases $x$ and $x'$ which are normatively identical, i.e. for all $\alpha \in A$, $E[v|x, \alpha] = E[v|x', \alpha]$. The proposition implies that whenever $x$ and $x'$ are evaluated together ($x, x' \in \mathbf{x}$) then they will be judged to have the same value, though this may not hold when they are evaluated apart.[9]

The same principle applies for choices, where the choice set is interpreted as the situation ($\mathbf{x}$): decisions may violate normative principles collectively, i.e. they

[8] I.e., $E_1^i = E[v|s(x^i), \alpha]$ and $E_P^i = E[v|x^i, \alpha]$.

[9] This holds because the restriction $E[v|x, \alpha] = E[v|x', \alpha]$ is *convex*: i.e. if it holds for every $\alpha$, then it also holds for an expectation taken using some beliefs $g(\alpha)$. This is true for most normative restrictions of interest, e.g. indifference ($\forall \alpha \in A$, $E[v|x, \alpha] = E[v|x', \alpha]$), dominance ($\forall \alpha \in A$, $E[v|x, \alpha] < E[v|x', \alpha]$), or separability (if $X = \mathbb{R}^2$, then $\forall \alpha \in A, x \in X, a, b \in \mathbb{R}$, $E[v|(x_1, x_2), \alpha] - E[v|(x_1 + a, x_2), \alpha] = E[v|(x_1, x_2 + b)] - E[v|(x_1 + a, x_2 + b), \alpha]$).

cannot be rationalized by any beliefs over $\alpha$, but every individual choice can be rationalized:

*Remark* 5. $\forall \alpha \in A$, $\forall \mathbf{x} \in X^m$, $\exists g \in \mathbb{R}^A$, $\int_{\bar{\alpha} \in A} g(\bar{\alpha}) d\bar{\alpha} = 1$, such that

$$\arg\max_{x \in \mathbf{x}} \mathbf{E}_2^{\alpha} = \arg\max_{x \in \mathbf{x}} \int_{\bar{\alpha} \in A} E[v|x, \bar{\alpha}] g(\bar{\alpha}) d\bar{\alpha},$$

where $\mathbf{E}_2^{\alpha} = E[\mathbf{v}|\mathbf{x}, E[v|\mathbf{s}(\mathbf{x}), \alpha]]$.

This implies that decision-makers will never make a dominated choice, i.e. a choice which cannot be rationalized by any beliefs over $\alpha$.

Other restrictions on utility such as separability also have testable implications for decision-making. For example expected utility, when combined with a dominance restriction, implies that the decision-maker will never make a stochastically dominated choice. Even without dominance, separability implies that no-one would choose, for example, a lottery between either a banana or an apple, when they also could choose an apple for sure, or a banana for sure.

## 1.4 Influence of Irrelevant Associations

In this section I show the connection between biases and real-world correlations: in short, when judgment is positively influenced by an irrelevant feature, then that feature must be associated with higher values in the real world at a *superficial* level, i.e. it is associated with higher values under a coarsening of the world, $s(x)$.

A decision-maker who forms judgments in this 2-stage manner will appear to be making judgments inattentively, or using heuristics: i.e., they will be influenced by features that are only superficially relevant. This is a common observation in empirical literature on judgment biases - that biases reflect superficial correlations - discussed in a later section.

Formally, consider some pair of cases $x$, $x'$, which are normatively identical - i.e., have the same expected value for all $\alpha$ - but which are superficially different - $s(x) \neq s(x')$. Because they are superficially different, System 2 may evaluate them differently ($E_2' \neq E_2$). We wish to state conditions under which System 2's bias goes in the same direction as System 1's bias, i.e. where:

$$\forall \alpha \in A,\ E_1' > E_1 \implies E_2' \geq E_2.$$

This can be interpreted as the bias going in the same direction as associations in the world, because $E_1' - E_1 = E[v|s(x'), \alpha] - E[v|s(x), \alpha]$, which is an objective fact about the true distribution of values in the world. For example, it could represent the fact that alternatives which are written in red ink tend to be better choices (and therefore explain a bias, if decision-makers are more likely to choose an alternative written in red ink, even when they know the color of the ink is uninformative).

I show that this conclusion will hold under two regularity conditions. First, if System 2 always regards an increase in System 1's signal as good news, no matter what System 2's private information is; I call this "congruence." Second, if the two cases, $x$ and $x'$, induce a symmetric uncertainty about $\alpha$, such that $E_1$ and $E_1'$ are treated in the same way in updating $E_p$. I discuss examples where these assumptions do not hold below.

**Proposition 2** (Influence of Irrelevant Associations)**.** *For any $x, x' \in X$, such that $\forall \alpha \in A$, $E[v|x, \alpha] = E[v|x', \alpha]$, if*

*(i) $f$ is congruent, i.e. for any $x'' \in X$,$\alpha, \alpha' \in A$:*

$$E[v|s(x''), \alpha'] > E[v|s(x''), \alpha] \implies E[v|x, E[v|s(x''), \alpha']] \geq E[v|x, E[v|s(x''), \alpha]],$$

*and (ii) $E_1$ and $E_P$ have the same relationship under $x$ and $x'$:*

$$f(E_1, E_P|x) = f(E_1, E_P|x'),$$

*then*

$$\forall \alpha \in A,\ E_1' > E_1 \implies E_2' \geq E_2,$$

*where $E_1 = E[v|s(x), \alpha]$, $E_1' = E[v|s(x'), \alpha]$, $E_2 = E[v|x, E[v|s(x), \alpha]]$, and $E_2' = E[v|x', E[v|s(x'), \alpha]]$.*

The congruence assumption can fail when System 2 holds information that reverses the significance of $E_1$. In terms of the Gaussian model below this would occur when, for some $i$, $\hat{x}_i < 0$, i.e. when System 2 believes that attribute 2 has the *opposite* relationship to $v$ than it usually has.

The symmetry assumption can fail when the cases induce different levels of uncertainty: suppose that System 2 is less certain about $\alpha$ under $x'$ than under

$x$, and therefore regards $E_1'$ as less informative than $E_1$. Then under $x'$ System 2 will use relatively less of System 1's signal, and relatively more of his prior, meaning that it's possible that $E_1' > E_1$ but $E_2' < E_2$. In the Gaussian model discussed below this would occur if, for example, $\bar{x} = (1, 1, 0)$, $\bar{x}' = (1, 0, 1)$, with $\sigma_2^2 > \sigma_3^2$, then when $\hat{x} = \hat{x}' = (1, 0, 0)$, it will be possible that $E_1' > E_1$ but $E_2' < E_2$, because in each of the two cases $E_1$ will be discounted by a different factor.

# 2 Gaussian Model

In this section I represent each case $x$ as a vector of attributes, and show that this allows for a richer characterization of biases.

Under the assumed functional form, System 2's problem can be seen as *re-weighting a weighted average*. I let $x$ be defined as a pair of vectors, $x = (\bar{x}, \hat{x})$, with $\bar{x}, \hat{x} \in \mathbb{R}^n$, where $\bar{x}$ represents the low-level or superficial features of $x$ (i.e., $s(x)$), and $\hat{x}$ represents high-level features. System 1 weights the superficial features of the case, $\bar{x}_1, \ldots \bar{x}_n$ according to its private information about associations $(\alpha_1, \ldots, \alpha_n)$. System 2 wishes to adjust the weighting using $(\hat{x}_1, \ldots, \hat{x}_n)$, but cannot perfectly infer $\alpha$, and so System 2's estimate, $E_2$, will be biased relative to the pooled-information benchmark; in particular, it will respond to changes in $\bar{x}$ which are known to be irrelevant, and the responses will be in the direction of empirical associations $(\alpha)$.

I assume that the pooled-information expectation is multiplicative in each element:

$$E_P = E[v|\alpha, x] = \sum_{j=1}^{n} \alpha_j \bar{x}_j \hat{x}_j.$$

Intuitively, a case consists of $n$ different attributes $(\bar{x}_1, \ldots, \bar{x}_n)$; in most applications these can be thought of as dummy variables, with $\bar{x}_i \in \{-1, 1\}$.[10] Each attribute $\bar{x}_i$ has an association with $v$, equal to $\alpha_i$, known only to System 1 (for example, each $\alpha_i$ could represent the positive or negative associations of different features of an apartment). However in each case System 2 observes information which modifies the weight appropriate to each attribute, represented by $\hat{x}_i$: for

[10] It is convenient to use $\{-1, 1\}$ instead of $\{0, 1\}$, so that both realizations have the same variance.

example, if System 2 learns that attribute $i$ is irrelevant in the current case, this is represented by learning that $\hat{x}_i = 0$.

I assume that $\alpha$, $\bar{x}$ and $\hat{x}$ are independently distributed, i.e.,

$$f_m(\mathbf{v}, \mathbf{x}, \alpha) = \left( \prod_{i=1}^{m} f(v^i | x^i, \alpha) f(\hat{x}^i) f(\bar{x}^i) \right) f(\alpha).$$

When there are $m$ cases the expectations can be expressed in matrix form:

$$\begin{aligned}
\underbrace{E_P}_{m\times 1} &= E[v|\bar{X}, \hat{X}, \alpha] = (\underbrace{\bar{X}}_{m\times n} \circ \underbrace{\hat{X}}_{m\times n}) \underbrace{\alpha}_{n\times 1} \\
E_1 &= E[v|\bar{X}, \alpha] = (\bar{X} \circ E[\hat{X}])\alpha \\
E_2 &= E[v|E_1, \bar{X}, \hat{X}] = (\bar{X} \circ \hat{X}) E[\alpha|E_1, \bar{X}],
\end{aligned}$$

where $\circ$ is the element-wise product of two matrices ($(P \circ Q)^i_j = P^i_j Q^i_j$), with $P^i_j$ referring to attribute $j$ of case $i$. Finally I assume that $\alpha$ is distributed normally,

$$\alpha \sim N(E[\alpha], \Omega),$$

with a covariance matrix $\Omega$.

In the following derivations I normalize $E[\hat{X}^i_j] = 1$ for all $i, j$. Because $\alpha$ is Gaussian, $E_1$ and $\alpha$ will have the following joint distribution:

$$\begin{pmatrix} E_1 \\ \alpha \end{pmatrix} \sim N\left( \begin{pmatrix} \bar{X}E[\alpha] \\ E[\alpha] \end{pmatrix}, \begin{pmatrix} \bar{X}\Omega\bar{X}' & \bar{X}\Omega \\ \Omega\bar{X}' & \Omega \end{pmatrix} \right),$$

implying that the conditional expectation of $\alpha$ can be expressed with a Schur decomposition:

$$E[\alpha|\bar{X}, E_1] = E[\alpha] + \Omega\bar{X}'(\bar{X}\Omega\bar{X}')^{-1}(E_1 - \bar{X}E[\alpha]).$$

**Proposition 3.** *System 2's expectation and bias will be:*

$$\begin{aligned}
E_2 &= (\bar{X} \circ \hat{X})E[\alpha] + (\bar{X} \circ \hat{X})\Omega\bar{X}'(\bar{X}\Omega\bar{X}')^{-1}\bar{X}(\alpha - E[\alpha]), \\
E_2 - E_P &= (\bar{X} \circ \hat{X}) \left( \Omega\bar{X}'(\bar{X}\Omega\bar{X}')^{-1}\bar{X} - I_n \right) (\alpha - E[\alpha]),
\end{aligned}$$

*where $I_n$ is an $n \times n$ identity matrix.*

We can make the following observations regarding the bias:

**(1) Bias will only occur when both systems have unexpected information.** In other words, judgment will be unbiased if either $\alpha = E[\alpha]$ or $\hat{X} = E[\hat{X}]$.[11]

**(2) Bias will disappear with sufficiently many cases.** If there are as many observations as cases ($n \geq m$), and $\bar{X}$ is of full rank (i.e., the cases are not collinear), then there will be no bias, because $\alpha$ will be exactly identified by $E_1$.[12]

**(3) Some situations can partially identify $\alpha$.** In situations where the full vector $\alpha$ is not identified, individual elements may be identified. An element $\alpha_i$ will be exactly identified ($E[\alpha_i|\bar{X}, E_1] = \alpha_i$) if the $i$th column of $\bar{X}$ is independent of the other columns, i.e. if it has a non-zero residual when projected on the other columns. A sufficient condition for this will be if some pair of cases differ only in dimension $i$, i.e. for some $j$, $k$:

$$\bar{X}_i^j \neq \bar{X}_i^k, \text{ and } \bar{X}_{-i}^j = \bar{X}_{-i}^k.$$

In other words, if two cases differ only in dimension $i$, this isolates the effect of dimension $i$ and we can exactly infer its contribution to our judgment.[13]

It is useful to examine the situation where we have a single case ($m = 1$). From this point I will assume that the covariance matrix $\Omega$ is diagonal with elements $\sigma_1, \ldots, \sigma_n$.

**Corollary 2.** *When $m = 1$ then*

$$\begin{aligned} E_2 &= \sum_{i=1}^{n} \bar{x}_i \hat{x}_i E[\alpha_i] + \hat{\phi} \left( \sum_{k=1}^{n} \bar{x}_k (\alpha_k - E[\alpha_k]) \right) \\ E_2 - E_P &= \sum_{k=1}^{n} \bar{x}_k (\alpha_k - E[\alpha_k]) \left( \hat{\phi} - \hat{x}_k \right), \end{aligned}$$

[11] Judgment will also be unbiased if $\hat{x}_i^j$ is constant for all $i, j$, i.e. if the second agent wishes to discount all dimensions by an equal amount. To see this, start with the expression for $E_2 - E_P$ and substitute in $\hat{X} = k1_{n\times m}$, where $k$ is a real number, and $1_{n\times m}$ is an $n \times m$ matrix of 1s, causing $E_2 = E_P$.

[12] If $\bar{X}$ has full rank then $\bar{X}'\bar{X}$ is invertible, and we can therefore show that $E[\alpha|E_1] = \alpha$ by prepending $(\bar{X}'\bar{X})^{-1}\bar{X}'\bar{X}$ into the expression for $E[\alpha|\bar{X}, E_1]$ and substituting $E_1 = \bar{X}\alpha$.

[13] This implies that bias will be zero if $i$ is the only unusual dimension ($\hat{x}_j = 1,\ \forall j \neq i$).

*where*

$$\hat{\phi} = \left( \sum_{i=1}^{n} \hat{x}_i \frac{\bar{x}_i^2 \sigma_i^2}{\sum_{j=1}^{n} \bar{x}_j^2 \sigma_j^2} \right).$$

The term $\hat{\phi}$ is a weighted average of the elements of $\hat{x}$, and summarizes how much weight System 2 gives to $E_1$. The overall bias can be thought of as a mixture of over-weighting of some dimensions (where $\hat{\phi} > \hat{x}_k$), and underweighting others (where $\hat{\phi} < \hat{x}_k$).

**(1) The direction of the bias is predictable.** It is simplest to consider deviations in a single dimension $k$, i.e. by conditioning on $\alpha_k$, $\hat{x}_k$:[14]

$$E[E_2 - E_P | \alpha_k, \hat{x}_k] = -\bar{x}_k(\alpha_k - E[\alpha_k])(\hat{x}_k - 1) \left( 1 - \frac{\bar{x}_k^2 \sigma_k^2}{\sum_{j=1}^{n} \bar{x}_j^2 \sigma_j^2} \right).$$

From this expression we can read off the effect of the bias. The final term in brackets will be between 0 and 1; it represents the share of variance in $E_1$ due to $\bar{x}_k$, and therefore the amount of adjustment that should be applied. Suppose $\bar{x}_k > 0$, i.e. $x$ has feature $k$, then the bias will be positive if $\alpha_k > E[\alpha_k]$ and $0 < \hat{x}_k < 1$. Intuitively, judgment will be biased upwards in this case because (i) System 2 is trying to diminish the effect of the attribute, because $\hat{x}_k < 1$, but (ii) it under-corrects for the influence, because the association is more positive than expected ($\alpha_k > E[\alpha_k]$).

**(2) The size of the bias is non-monotonic in $\bar{x}$.** The expression above also shows that, as $\bar{x}_k$ increases, the sign of the bias remains the same, but the magnitude is non-monotonic: the bias will be proportional to $\frac{\bar{x}_k}{\kappa + \bar{x}_k^2}$, i.e. it will be increasing then decreasing, as in Figure 2. Intuitively, as $\bar{x}_k$ gets larger, $E_1$ becomes less informative about the $\alpha$ parameters of interest, and System 2 applies a larger discount to $E_1$, asymptotically ignoring $E_1$ and using only its priors over $\alpha$.

Finally we discuss the solution where there are exactly two cases. Here we can directly discuss the effect of comparisons on choice.

**Corollary 3.** *When $m = 2$, with cases $x = (\bar{x}, \hat{x})$ and $x' = (\bar{x}', \hat{x}')$, then for case*

[14] I.e., setting $\alpha_j = E[\alpha_j] = 0$, $\hat{x}_j = E[\hat{x}_j] = 1$ for $j \neq k$.

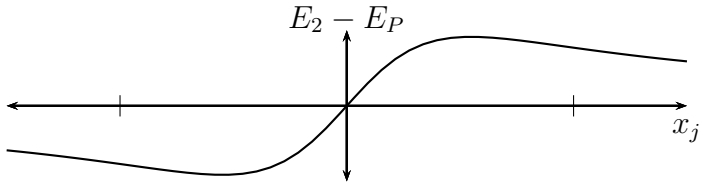



**Figure 2:** The effect of $x_j$ on System 2's bias, $E_2 - E_P$, when $\hat{x}_j = 0$ and $\alpha_j > 0$.

$(\bar{x}, \hat{x})$*:*

$$E_2 - E_P = \frac{\Sigma_{\bar{x}^2\hat{x}\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - \Sigma_{\bar{x}\bar{x}'\hat{x}\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2}}{\Sigma_{\bar{x}^2\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - \left(\Sigma_{\bar{x}\bar{x}'\sigma^2}\right)^2}\Sigma_{\bar{x}(\alpha - E[\alpha])} - \frac{\Sigma_{\bar{x}^2\hat{x}\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2} - \Sigma_{\bar{x}\bar{x}'\hat{x}\sigma^2}\Sigma_{\bar{x}^2\sigma^2}}{\Sigma_{\bar{x}^2\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - \left(\Sigma_{\bar{x}\bar{x}'\sigma^2}\right)^2}\Sigma_{\bar{x}'(\alpha - E[\alpha])} - \Sigma_{\bar{x}\hat{x}(\alpha - E[\alpha])},$$

*where* $\Sigma_{ab} = \sum_{i=1}^{n} a_i b_i$ *for any* $n$*-vectors* $a, b$*, etc.*

We can make some quick observations using this formula:

**(1) If $x' = x$, then addition of $x'$ does not affect the bias.**

**(2) Biases and implicit knowledge can be inferred from joint and separate evaluations.** Suppose that the two cases differ in only one superficial respect, i.e. for some $i$ let $\bar{x}'_i = 1$, $\bar{x}_i = -1$,[15] and $\bar{x}_j = \bar{x}'_j$ for $j \neq i$ (and $\hat{x} = \hat{x}'$), and suppose that we observe $E_2$ for each case evaluated separately, as well as $E_2$ for the two cases evaluated jointly (denote the four evaluations as $E_2^S, E_2'^S, E_2^J, E_2'^J$). From the joint evaluation we can directly infer $\alpha_i \hat{x}_i = \frac{1}{2}(E_2'^J - E_2^J)$. The difference between the two separate evaluations will be equal to:

$$E_2'^S - E_2^S \;\; = \;\; \hat{x}_i 2E[\alpha_i] + \hat{\phi} 2 \left(\alpha_i - E[\alpha_i]\right),$$

where $\hat{\phi}$ is defined in corollary 2, which represents the discount factor which System 2 applies to $E_1$; it is a weighted average of all the components of $\hat{x}$.

[15] Assuming $x_i \in \{-1, 1\}$ is shorthand for having pairs of dummy variables with equal variance. Suppose that $x$ is partitioned into pairs of dummy variables, such that $x_i \in \{0, 1\}$, and for every odd $i$, $x_i + x_{i+1} = 1$ and $\sigma_i^2 = \sigma_{i+1}^2$. These pairs could be interpreted as binary variables: e.g. male/female, yellow/blue. To save space I simply define a representation $\bar{x}$ with half as many dimensions, where $\bar{x}_k = x_{2k-1} - x_{2k}$, implying $\bar{x}_k \in \{-1, 1\}$, and $\bar{\sigma}_k^2 = 2\sigma_{2k}^2$.

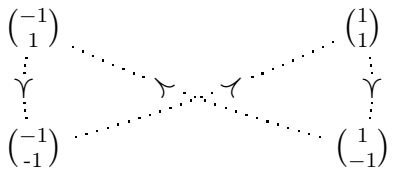
It follows that the difference between joint and separate evaluations reveals the direction of the bias:

$$\left(E_2'^J - E_2^J\right) - \left(E_2'^S - E_2^S\right) = 2(\hat{x}_i - \hat{\phi})(\alpha_i - E[\alpha_i]).$$

If the difference between joint and separate evaluation is positive, then either $\hat{x}_i > \hat{\phi}$ and $\alpha_i > E[\alpha_i]$, or the reverse. In some cases it may be reasonable to assume that $\hat{x}_i < \hat{\phi}$, meaning that we think the decision-maker wishes to discount dimension $i$ more than the other dimensions. Then the difference between joint and separate evaluation directly tells us the difference between the implicit and explicit knowledge: i.e. whether $\alpha_i > E[\alpha_i]$ or $\alpha_i < E[\alpha_i]$.

More concretely, suppose we observe that judgments are influenced by some feature, e.g. that auction bids are influenced by prior selling price, or asset prices are influenced by ticker symbol, or judges' verdicts are influenced by the name of the defendant. This effect could either be an explicit preference, or an implicit bias. The model in this paper tells us that it is a bias if the difference in evaluation shrinks in joint evaluation, i.e. when two cases are compared which differ only in that respect.

**(3) Biases and implicit knowledge can be inferred from from intransitive choices:** Suppose we consider cases which differ only along two dimensions, $i$ and $j$, with $\bar{x}_i, \bar{x}_j \in \{-1, 1\}$, and denote each case as $(\bar{x}_i, \bar{x}_j)$. Now suppose that we observe the following pattern of choices within pairs of cases (here $a \succ b$ is used to denote the choice of $\{a\}$ from the choice set $\{a, b\}$):

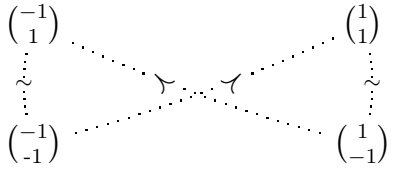



These choices are intransitive: whenever two cases are compared which differ only in dimension $j$, then the higher $\bar{x}_j$ wins, but when the cases also differ in the other dimension, then the lower $\bar{x}_j$ always wins. We could call this an *explicit* positive preference for attribute $j$, but an *implicit* negative preference for $j$.

Formally, we can infer from the two Southbound arrows that $\hat{x}_j \alpha_j > 0$. From the North-East arrow we can infer $\hat{x}_i \alpha_i^{NE} + \hat{x}_j \alpha_j^{NE} < 0$, where $\alpha_k^{NE} = E[\alpha_k] +$

$\frac{\sigma_k^2}{\sigma_i^2+\sigma_j^2}(\alpha_i+\alpha_j)$ for $k \in \{i,j\}$, and for the North-West arrow we can infer $\hat{x}_i\alpha_i^{NW} - \hat{x}_j\alpha_j^{NW} > 0$, where $\alpha_k^{NW} = E[\alpha_k] + \frac{\sigma_k^2}{\sigma_i^2+\sigma_j^2}(\alpha_k - \alpha_{-k})$. Combined, these imply that the bias is positive (Northbound), i.e. $-(\hat{x}_j - 1)(\alpha_j - E[\alpha_j]) > 0$, and therefore either $\hat{x}_j < 1$ and $\alpha_j > E[\alpha_j]$, or $\hat{x}_j > 1$ and $\alpha_j < E[\alpha_j]$. If there is reason to believe that the subject is less confident than usual about about dimension $j$, i.e. $\hat{x}_j < 1$, then we can conclude the subject has implicit positive knowledge about $\bar{x}_j$, i.e. the association between $\bar{x}_j$ and $v$ is stronger than the subject is consciously aware.

We can make stronger inferences if we observe indifferences:



These choices imply that $\hat{x}_j = 0$, so $\alpha_j > E[\alpha_j]$, i.e. the association between $\bar{x}_j$ and $v$ must be stronger than System 2 realizes. An example was given in the introduction: if a decision-maker is indifferent about weather in direct comparisons, but prefers apartments viewed on sunny days in indirect comparisons, it implies they have positive implicit associations with the sun, but are consciously aware that the weather is irrelevant in the current situation.

# 3 Testing the Model

## 3.1 Evidence for Internal Consistency

In this section I discuss existing available evidence for Remark 4, i.e. that normative restrictions on preferences are violated only *indirectly*.

This is not true without qualification. Every incorrect answer to a question on a mathematics test is a dominated choice. The important point for this model is that *between*-choice violations of normative rules occur even with with simple outcomes, and high stakes, while *within*-choice violations are observed only under much more demanding conditions, i.e. when outcomes are reasonably

complicated, or when stakes are low.[16]

An alternative interpretation of the evidence discussed in this section is that people use a "consistency heuristic" or have a "taste for consistency," or use an "editing rule" (Kahneman and Tversky (1979)).[17] In fact the model in this paper can be interpreted as an explanation of why people act as if using such a heuristic: because by averaging their intuitive judgments, they can predict value more accurately. It would be a challenge to this model if the consistency heuristic was only obeyed when decision-making is observed by an external observer.

I discuss (1) evidence for a low rate of dominance violations in laboratory choices; and (2) examples of biases where irrelevance and dominance constraints are violated indirectly, but not directly.

**Infrequency of dominated choice.** The literature on laboratory choice over gambles has found that, although between-subject and between-choice inconsistencies are common, it is very rare for subjects to directly choose a dominated alternative. Carbone and Hey (1995) say "[w]hat is startling ... are the results of the satisfaction or violation of dominance ... [with a] mean violation rate of just 0.3 percent. In contrast the average inconsistency rate of the repeated pairs was 12 percent." Similar findings are discussed in Loomes and Sugden (1998) and Hey (2001).

**Different valuations for equivalent alternatives.** There exist a number of well-known biases in which irrelevant features affect judgment in between-subject experiments: for example anchoring effects, price range effects, and loss/gain framing effects.[18] The model of implicit knowledge predicts that the effects will disappear in joint evaluation. Unfortunately papers which report re-

[16] People do sometimes directly choose a stochastically dominated gamble, but most example in the literature involve choices among gambles with at least four different outcomes (Tversky and Kahneman (1986), Birnbaum et al. (1999)). Caplin and Dean (2013) find mistakes when subjects have to count a large number of dots. In experiments documenting "bracketing", subjects make dominated choices when they have to combine pairs of gambles Tversky and Kahneman (1981), Rabin and Weizsacker (2009).

[17] A number of experimental papers find within-subject question order effects, which they interpret as people modifying later answers to be consistent with earlier answers (Falk and Zimmermann (2013)). The model in this paper predicts history-dependence, but not order-dependence (i.e., the set of prior questions does matter, but their order does not).

[18] Anchoring experiments find that willingness to pay is strongly affected by an arbitrary anchor (Fudenberg et al. (2010), among others):

$$WTP(\text{high anchor}) > WTP(\text{low anchor})$$

sults from joint evaluation are rare. A notable exception is Mazar et al. (2013): they find that, when presented separately, willingness to pay (WTP) judgments are sensitive to the modal price in the price distribution. Normatively, the modal price is irrelevant, because WTP is elicited with the BDM mechanism, and the price distribution is chosen by a public coin flip. They find that:

$$WTP(\text{high modal price}) > WTP(\text{low modal price}).$$

Mazar et al. additionally ask subjects to state WTP judgments when both price distributions are presented at once (I denote this as $WTP_{JOINT}(\cdot)$), and they find that the difference almost disappears, i.e.:

$$WTP_{JOINT}(\text{high modal price}) \simeq WTP_{JOINT}(\text{low modal price}),$$

as predicted by the model of implicit knowledge.

**Higher valuation for dominated alternative.** There exist a number of well-known biases in which subjects indirectly evaluate a strictly dominated prospect as better than a dominating prospect (i.e. a pattern $p \succ q$, $q \succ p'$, where $p'$ dominates $p$). In most cases the experimenters never test whether subjects *directly* choose a dominated alternative ($p \succ p'$), presumably because they doubt it would occur.[19]

Starmer (1999), for example, documents that a prospect is valued more highly

Beauchamp et al. (2012) find that expressed WTP judgments for gambles are sensitive to the maximum value available in the price list:

$$WTP(\text{high max value}) > WTP(\text{low max value})$$

Tversky and Kahneman (1981) ask subjects to choose between strategies, given that 600 people are at risk from some disease, and find the median response differs depending on the framing:

$$\begin{aligned} (\text{200 saved}) &\succ \begin{pmatrix} \text{600 saved}, \frac{1}{3} \\ \text{0 saved}, \frac{2}{3} \end{pmatrix} \\ (\text{400 die}) &\prec \begin{pmatrix} \text{0 die}, \frac{1}{3} \\ \text{600 die}, \frac{2}{3} \end{pmatrix} \end{aligned}$$

[19] Gneezy et al. (2006) say, to explain why they did not perform a within-subjects treatment, that "[dominance] is so normatively compelling that it will be followed when the problem presentation suggests that the axiom should be applied."

when low-probability gains are split:[20]

$$\begin{aligned}(£14,\ 20\%) &\succ (£8,\ 30\%)\\ (£14,\ 20\%) &\prec \begin{pmatrix} £8,\ 15\% \\ £7.75,\ 15\% \end{pmatrix}.\end{aligned}$$

Andreoni and Sprenger (2011) find that prospects are valued more highly when the outcome is certain:[21]

$$\begin{aligned}(\$10,\ 100\%) &\succ (\$30,\ 50\%)\\ \begin{pmatrix} \$10,\ 95\% \\ \$30,\ 5\% \end{pmatrix} &\prec (\$30,\ 50\%).\end{aligned}$$

Gneezy et al. (2006) find a similar effect in WTP for book vouchers:[22]

$$WTP(\$100 \text{ voucher}, 100\%) > WTP \begin{pmatrix} \$100 \text{ voucher},\ 90\% \\ \$200 \text{ voucher},\ 10\% \end{pmatrix}.$$

Birnbaum (1992) finds that adding a low-value prize lowers WTP for a gamble:

$$WTP \begin{pmatrix} \$96, 90\% \\ \$0, 10\% \end{pmatrix} > WTP \begin{pmatrix} \$96, 90\% \\ \$20, 10\% \end{pmatrix}.$$

List (2002) finds a similar pattern in bids for different sets of baseball cards: adding bad cards lowers valuations:

$$bid\,(10 \text{ good cards}) > bid \begin{pmatrix} 10 \text{ good cards} \\ +3 \text{ bad cards} \end{pmatrix}.$$

Finally Hsee (1998) finds that increasing the size of an ice cream cup lowers WTP

[20] These choices are consistent with first-generation prospect theory, but not cumulative prospect theory.

[21] This paper uses within-subject data: however subjects answer dozens of questions, so the dominance relation is not transparent.

[22] They also find the effect in choice data. The effect is replicated in Simonsohn (2009), but not Rydval et al. (2009) or Keren and Willemsen (2009).

for ice cream:

$$WTP\begin{pmatrix}\text{7oz ice cream}\\ \text{in 5oz cup}\end{pmatrix} > WTP\begin{pmatrix}\text{8oz ice cream}\\ \text{in 10oz cup}\end{pmatrix}.$$

Hsee's paper also elicits WTP in joint evaluation, which causes the rankings to reverse.

In each of these cases normative dominance constraints are indirectly violated. This could be predicted by a model of inattention: e.g., if certain superficial characteristics were used as proxies of value (such as *being risky*, or *being all good cards*, or *being less-than-full*). However such an explanation would also predict that dominance constraints will be directly violated, a fact that does not seem to be true.

## 3.2 Evidence for Associative Judgments

The model predicts that, when judgment is found to be influenced by an irrelevant feature or frame, then the direction of the influence will correspond to the real-world empirical association between that feature and the underlying value. The prediction is difficult to test because it requires knowledge of the true real-world correlations. This difficulty is common to all theories of inattentive judgment: predicting the direction of a bias requires knowing the true frequency distribution of outcomes.[23]

Nevertheless this argument is frequently made on a case by case basis. Gigerenzer and Gaissmaier (2011) argue that many biases in judgment are due to using heuristics that are "ecologically rational". Rand et al. (2013) state that many findings in the interpersonal preference literature can be explained as the application of heuristics that are *ordinarily* rational, although not rational in the current situation. Erev and Haruvy (2013) state that many puzzling features in the "choice from experience" paradigm can be explained by appeal to subjects using rules which are *usually* optimal.

A good test of this class of models would be to show that interpersonal differ-

[23] A similar issue arises in the literature on Bayesian perception and cognition: to judge whether an inference is optimal requires assumptions on what are appropriate priors (Jones and Love (2011)).

ences in biases correspond to differences in experience. Peysakhovich and Rand (2013) find that players are more likely to cooperate if cooperation has been a good strategy in the past (and this holds even when it is transparent that their experience is not informative about the current returns to cooperation). Erev and Haruvy (2013) discuss evidence that recent experience influences choice in a similar way. Caplin and Martin (2012) show in an experiment that subjects' choices are influenced by uninformative labeling of an alternative as the "default alternative", when in prior experience the default has been correlated with higher payoffs.

An ideal test of this aspect of the model would have the following two-stage structure. First, expose subjects to a series of cases in which some attribute $\bar{x}_i$ (e.g., color) is either negatively or positively associated with value. Second, elicit judgments in new cases where $\bar{x}_i$ is known by the subjects to be uninformative about value (i.e., set $\hat{x}_i = 0$). The model of implicit knowledge makes two predictions. First, judgment will be affected by $\bar{x}_i$ , despite its irrelevance, in the direction of the learned association: i.e., when subjects attempt to ignore the attribute they will be unsuccessful. Second, when subjects simultaneously evaluate two cases which differ only in $\bar{x}_i$, then the two cases will be judged to have equal value.[24] I.e., the bias will not be detectable in joint evaluation.

# 4 Discussion

## 4.1 Related Literature in Psychology

The distinction between separate systems, with access to different sources of information, has often been made in psychology.

In the study of perception it has become standard to distinguish between stages of processing, with the assumption that early stages do not have access to high-level information: they are "informationally encapsulated"; this assumption is also called "modularity" or "bottom-up" processing (Marr (1982), Fodor (1983),

[24]This is related to the literature in psychology on implicit learning, e.g. Reber (1989). An important difference is that in those experiments the implicitness is determined by whether subjects can give verbal reports of the knowledge, and the evidence for this is controversial (Newell and Shanks (forthcoming)). An advantage of my design is that the existence of implicit knowledge can be identified entirely from choice behavior.

Pylyshyn (1999), Feldman (2013)). This assumption fits neurological evidence of a hierarchical organization of brain regions, but it is also used to explain perceptual illusions. When looking out a train window, it sometimes appears that the platform is moving. This can be explained as our perceptual system making inferences without taking into account high-level knowledge, i.e. the prior knowledge that it is far more likely that the train is in motion, rather than the platform.[25] Implicit in this literature is the additional assumption that the early stages have their own private information: if our conscious brain had access to all the information used in interpreting perceptual data, then we could avoid biases just by thinking hard enough. In the case of the train platform this is true, we may have a feeling that the platform is moving but we know that it is really stationary. However this is not the case for many other illusions: often we are forced to trust the interpretations made by our perceptual systems, and we therefore make systematic mistakes (i.e., we are influenced by information known to be irrelevant) even when we concentrate on the problem. For example, the Müller-Lyer optical illusion, in which irrelevant bars affect judgment of the length of a line, remains even with a long inspection time (Predebon et al. (1998)).[26] These observations point towards the existence of a perceptual system that does not have access to high-level information (such as the belief that trains don't move), but, in addition, a conscious system that does not fully understand how the perceptual system works, and therefore cannot override its biases.

Two-system models are also popular in the study of judgment and decision-making in psychology (Sloman (1996), Stanovich and West (2000), Evans (2008), Kahneman (2011)); an important argument being that biases often appear to be optimal responses relative to the low-level information: Sloman (1996) says "[r]easoning performance is accurately predicted by judgments of similarity taken out of the problem context in the absence of any further assumptions about the knowledge that people bring to bear on the task." The first system is often

[25] Adelson (2000) discusses many elegant explanations of illusions using this framework.

[26] Additional evidence for $\alpha$ being implicit is the difficulty in teaching computers to perform simple perceptual tasks (Meer (2012)), i.e. translating our implicit knowledge of perceptual cues into explicit knowledge that can be used by computers. An extra complication in applying this model to perception is that System 1's private information, $\alpha$, may include case-specific information: i.e., perceptual data which is not consciously accessible. Existence of this information implies that the influence of irrelevant information may not disappear in joint evaluation, though it will decline.

described as having access to rich information about associations, but whether this information is private to that system (the crucial assumption in this paper) is not usually discussed.

A large literature exists on the existence of implicit knowledge, most famously Polanyi (1966). However laboratory tests of the hypothesis have a controversial history, and there is no agreed-upon criterion for demonstrating that knowledge is implicit (Newell and Shanks (forthcoming)). A sociological argument for the importance of implicit knowledge is given in Autor et al. (2014): he observes that computer scientists have made surprisingly slow progress in teaching computers to perform activities that they themselves find very simple (such as recognizing whether a sentence is grammatical; or recognizing images in a picture), and that recent progress has mainly used "machine learning," in which the machine teaches itself.

Tversky and Kahneman (1974), the paper which introduced the phrase "heuristics and biases," gives a motivating example from perception that fits the model in this paper very well. They note that people appear to systematically overestimate distances on foggy days, and that perceptual psychologists explain this as due to people rationally using the blurriness of an object as a cue for estimating its distance; therefore because fog makes everything blurrier, it makes everything seem more distant (Ross (1975)). However this effect is not rational when people are aware that it is a foggy day; in that case they should make an appropriate adjustment, and be influenced only by the *excess* blurriness of a given object, given the ambient fog. This point is discussed in Kahneman and Frederick (2005), where they assume that System 1 makes the inference without knowing that it is foggy, and they state that System 2 *could* make an optimal inference if it was activated.[27] They therefore predict that the bias will only exist when System 2 is not activated. In contrast the model in this paper shows that the bias will persist with implicit knowledge - i.e. if System 1's information about the correlation between blur and distance is not available to System 2.[28]

[27]"[a]lthough people are capable of consciously correcting their impressions of distance for the effects of ambient haze, they commonly fail to do so." On the other hand Kahneman (2011) states that "most of the work of associative thinking is silent, hidden from our conscious selves".

[28]In the Gaussian model, let $v$ be the distance of an object, let $x_i$ be its blurriness, let $\alpha_i > 0$ represent the positive relationship between blurriness and distance, and $E[\alpha_i] < \alpha_i$ represent that System 2 underestimates this relationship. Finally for a foggy day let $\hat{x}_i < E[\hat{x}_i]$

The model predicts that System 2 will *discount* visual impressions from System 1 on a foggy day, but because it does not know the relationship between blurriness and distance, a bias will persist. A small survey seems to support this interpretation:[29] of 40 subjects, 13% said "fog makes cars seem closer than they really are", 37% said "fog doesn't change perception of distance", and 50% said "fog makes cars seem farther away than they really are." This shows that at least half the sample underestimate the effect of fog on distance judgment, because the true effect is in fact very large: Cavallo et al. (2001) find "an average increase of 60% in the perceived distance of vehicles in fog as compared with normal visibility conditions."

## 4.2 Related Literature in Economics

The model in this paper shares a dual-agent structure with other recent models (Bodner and Prelec (2003), Bénabou and Tirole (2004), Fudenberg and Levine (2006), Brocas and Carrillo (2008)). However the results in those models are driven by the conflict of preferences, whereas in this paper the systems' preferences are aligned.[30] The two classes of model are complementary if, as seems likely, separate mental processes differ in both information and preferences.[31]

There are a number of models of imperfect memory which have similarities to the model in this paper. Many have the feature that memory is aggregated in some way that produces biases in behavior relative to an agent that has access to all their past experiences.[32] These models can explain decisions being affected by

to represent the fact that blurriness is less informative about distance than usual. Then the bias, $E_2 - E_P$, will be proportional to $-(\alpha_i - E[\alpha_i])(\hat{x}_i - E[\hat{x}_i])$, meaning that people will overestimate distances on days when it is foggy (when $\hat{x}_i < E[\hat{x}_i]$), and they will underestimate distances when it is clear ($\hat{x}_i > E[\hat{x}_i]$), both facts are confirmed in Ross (1975).

[29]The survey was run on Mechanical Turk, no demographic information was collected.

[30]In Fudenberg and Levine (2006) there is no asymmetry of information. In Brocas and Carrillo (2008) there is some asymmetry, but the short-run agent (the decision-maker) knows strictly more than the long-run agent, so if their preferences were aligned (as in this paper) there would be no bias. In Bodner and Prelec (2003) and Bénabou and Tirole (2004) one agent chooses strategically, to affect a second agent's beliefs about the first agent's preferences (i.e., self-image): information is asymmetric in only one direction, and preferences are effectively non-aligned, which is needed to generate a signaling motive.

[31]Brocas and Carrillo (2013) model two agents with aligned preferences, and one agent is only activated under certain conditions; however when both agents are activated all information is perfectly aggregated, so decisions will be unbiased.

[32]In Shapiro (2006) consumers are uncertain if they have positive memories about a product

irrelevant information, but only irrelevant information in the *past*, because the irrelevant information affects current behavior through being aggregated together with relevant information into coarse memories. These models do not predict judgment being influenced by *contemporaneous* irrelevant information, as in the framing effects that motivate this paper.

The model is also related to models of inattention (Sims (2005), Chetty et al. (2007), Caplin and Martin (2011), Woodford (2012), Gabaix (2012)): in the model of implicit knowledge, System 1's evaluations will be inattentive in the usual sense, because it receives only a coarse signal of $x$. The model in this paper could be thought of as a combination of limited-attention and limited-memory models: System 1 has limited attention, System 2 has limited memory. This allows it to reproduce the qualitative biases in limited attention models, but it can also explain why we rarely make dominated choices, as in a limited-memory or random preferences model.[33] The difference from inattention models can be thought of as whether inconsistencies are due to variation in perception of the choice set (which allows dominated choices), or variation in indifference curves (which rules them out).

Formally this paper is most similar to models of social learning (Chamley (2003)), in that each agent receives a private signal, one agent observes another agent's action, and we solve for the conditions under which information is efficiently aggregated. The model in this paper differs from the standard social learning setup in a few ways: in having a continuous action space (most herding models have a discrete action, to prevent the action being a sufficient statistic for the information); in having a many-to-one mapping between signal and the expectation of the underlying variable (which prevents $E_1$ from revealing $\alpha$); and in having just two agents, and thus being interested in perfect aggregation, not

because of a good experience, or because of seeing advertisements, and thus they may be influenced by uninformative advertising. In Baliga and Ely (2011) investors do not remember why they started a project, so may exhibit sensitivity to costs which are sunk. In Wilson (2002) agents have a limited number of states they can use in memory, so sometimes rationally ignore information, or are sensitive to the order in which information is received. Mullainathan (2002) and Schwartzstein (2012) are similar but additionally assume that agents are naive about how memories are retrieved.

[33]There is also a similarity to the "editing" phase of 1st generation prospect theory, proposed in Tversky and Kahneman (1979), in which dominated prospects are removed from the choice set.

aggregation in the limit. An important pair of related papers are Mueller-Frank and Arieli (2012) and Arieli and Mueller-Frank (2013), discussed earlier.[34]

## 4.3 Implications for Anomalies in Economic Decision-Making

The model of implicit knowledge can be used to interpret framing effects and related anomalies often found in economic decision-making.

This kind of effect can be explained within the framework of this paper if the feature that affects valuation is *usually* informative. For example some common biases can be rationalized if decoy alternatives, reference points, and price anchors are usually informative proxies for value. This case has been made explicitly in previous papers: Wernerfelt (1995), Prelec et al. (1997) and Kamenica (2008) argue that it can be rational for a decoy alternative to influence your decision; McKenzie and Nelson (2003) argue that it can be rational for a reference-point to influence your choice; Armstrong and Chen (2013) argue that it can be rational to be influenced by a reference price (or anchor). However each of these models predicts that if the feature is transparently randomized (and so made uninformative) then it should no longer have an effect on choice. Yet a variety of studies find framing effects which persist under explicit randomization (Fudenberg et al. (2010), Jahedi (2011), Mazar et al. (2013)). Thus these papers can be thought of as explaining the *direction* of bias for System 1; in turn the fact that the biases survive under randomization can be explained by System 2's unawareness of how System 1 uses these cues.

## 4.4 Economic Applications

***Marketing and Persuasion.*** The model in this paper can justify a role for marketing as creating *associations*, a common description used by practitioners, in contrast to informational or signaling roles usually used in economic models.[35]

[34]Example 2 in Mueller-Frank and Arieli (2012) fits the assumptions of the model in this paper: although agent 2 can infer agent 1's posterior over $v$, agent 2's posterior is not equal to the pooled-information posterior, because agent 2's private information is not conditionally independent of agent 1's private information.

[35]Iacobucci (2001), says "[t]he marketer's goal is to create and reinforce a cognitive network of associations, with positive attributes linked to their brand, and weak or negative associations linked to alternative brands."

In particular, judgments of value will be influenced by apparently irrelevant details (as documented in, for example, Bertrand et al. (2010)). Individuals who have superior awareness of implicit associations, $\alpha$, will have high economic value: they will be able to choose sets of characteristics, $x$, which evoke high valuations. Shapiro (2006) proposes a related model with the assumption that exposure to advertising regarding some product increases the probability that a consumer will recall a positive memory about that product. Uninformative advertising can therefore be effective when consumers do not recall the amount of advertising to which they were exposed. Shapiro's model predicts that uninformative advertising will only work with a lag; however with implicit knowledge, uninformative advertising can be persuasive immediately - a testable prediction.

***Within-Between Effects.*** There are many field studies which find irrelevant influences on decisions. The model in this paper predicts that irrelevant influences will be systematically larger in between-context studies than in within-context studies. For example, Beggs and Graddy (2009) find that, in art auctions, the prior selling price of a painting has a significant effect on current selling price (instrumented for by date of sale), and Brown et al. (2010) find that, in online auctions, the level of shipping price has a significant effect on the final gross selling price (i.e., the price paid including shipping). The model of implicit knowledge predicts that these effects will be greatly attenuated in within-situation studies: i.e., when subjects compare two similar paintings with different prior selling prices, or comparing two CDs for sale with similar total prices, but different division between base-price and shipping-price.

***Inferring Preferences from Choices.*** A recent topic of interest has been how to do welfare economics if decisions do not reflect utility-maximizing choices: Bernheim and Rangel (2009) propose that, although preferences cannot be exactly identified from choices, they could be bounded within the range of choices elicited under different frames. In practice these bounds may in fact be very large, and discovering the bounds requires knowledge of the full distribution of possible frames. The model presented in this paper has the advantage that welfare is well-defined, and exact solutions can, in principle, be inferred from decisions. In this model optimal decisions maximize $v$, and constrained-optimal decisions maximize $E_2$.

Here I note five implications for elicitation, and leave a fuller analysis to fu-

ture work: choices are more likely to reflect true preferences when (1) incentives are high; (2) people are in more familiar situations; (3) people are provided with a wider range of cases for consideration; (4) people are provided with comparisons which isolate the dimensions which are unusual; and (5) people are directly informed about empirical associations.

First, if System 1 is the default, and System 2 is only activated when there is sufficient time or incentives, then activating System 2 will reduce but not eliminate bias (i.e., $E_2 \neq E_P$). Larrick (2004) notes that incentives are often insufficient to debias judgments: "[t]here is little empirical evidence ... that incentives consistently improve mean decision performance ... incentives reduce biases in only a handful of cases."

Second, the Gaussian model predicts that bias will be lower when $\hat{x}$ is closer to $E[\hat{x}]$, i.e. when the high-level information is close to its average values. This could be interpreted as meaning that people tend to make better judgments and choices in "ordinary" or "familiar" situations.[36]

Third, as shown in Proposition 3, bias will tend to diminish when subjects are exposed to a larger set of cases.[37] There are some recent studies which argue the opposite - that people tend to make worse choices from larger choice sets (Iyengar and Kamenica (2010)) - though they propose an independent and complementary mechanism.

Fourth, the model predicts that not all comparisons are equal: certain comparisons can improve judgment more than others. The Gaussian model shows that, when only one dimension is unusual, then a comparison which isolates that dimension will perfectly debias judgment: for example, you may choose to visit the same apartment on both a sunny and a cloudy day, to isolate the effect of weather, and therefore debias your judgment. In experiments, if we are concerned that an irrelevant detail is affecting choice (i.e., a framing effect), then the model recommends simultaneously presenting subjects with multiple versions of the same case, varying the irrelevant detail, and therefore getting closer to the true preferences.

[36]This seems to be true for perception: for example, our ability to recognize faces is significantly worse when the faces are upside down (Sinha et al. (2006)); this is commonly taken as evidence for the encapsulation of visual processing.

[37]Larrick (2009) claims that judgments are generally improved by "broadening the decision frame", by considering multiple objectives, multiple alternatives and multiple outcomes.

Finally, the model predicts that biases could be eliminated if subjects were directly told the contents of $\alpha$: i.e., if they were consciously aware of the associations that their intuitions used. This advice is followed in dealing with visual illusions, e.g. an Airbus training manual warns "[f]lying in light rain, fog, haze, mist, smoke, dust, glare or darkness usually creates an illusion of being too high" (Airbus (2011)). In an economic context, decisions could be improved if.

# 5 Conclusion

This paper argues that many puzzles of human judgment can be explained by a simple model: when we form judgments we take advice from a separate dedicated system, which has superior information about prior experiences with similar cases, but which fails to take into account high-level or abstract information about the current case.

This model reconciles two facts about decision-making: (1) we commonly violate normative restrictions on behavior, as if we are influenced by irrelevant associations; but (2) it is rare that we *directly* violate restrictions on behavior.

There are a number of interesting issues raised that I leave for future work. One is why our brain should be structured in such a modular way. Another is a more precise study of how information is partitioned between systems, i.e. what information is "low level" and "high level". A third is which value, $v$, is inferred: from a given set of cues, there are potentially multiple different underlying values that System 1 could wish to infer.[38] Finally, in some cases associations may become hardened into preferences, such that an attribute becomes valued for its own sake, not just for its predictive value with regard to intrinsic rewards (means become ends); it may be interesting to extend the model in this direction.[39]

[38]The model of "attribute substitution" in Kahneman and Frederick (2005) could be rationalized in this way: the automatic system evaluates one attribute, and the reflective system uses that signal to infer a different attribute, causing a bias towards the attribute which is automatically inferred. For example, suppose the automatic system evaluates *representativeness* of an event, and when making probability judgments the reflective system uses that estimate to infer the *likelihood* of the event, causing an apparent "representativeness" bias in judgments of likelihood.

[39]In the neurological learning literature a common finding is that animals stop performing a rewarded action when the reward is devalued; however if the subjects are "over-trained" then the performance can persist (Daw et al. (2005)). A simple example of means becoming ends in

# References

**Adelson, Edward H**, "Lightness Perception and Lightness Illusions," in M. Gazzinaga, ed., *The New Cognitive Neurosciences*, The MIT Press, 2000, pp. 339–351.

**Airbus**, "Flight Operations Briefing Notes: Human Performance: Visual Illusions Awareness," Technical Report, Airbus 2011.

**Andreoni, James and Charles Sprenger**, "Uncertainty equivalents: Testing the limits of the independence axiom," Technical Report, National Bureau of Economic Research 2011.

**Arieli, Itai and Manuel Mueller-Frank**, "Inferring Beliefs from Actions," SSRN Working Paper 2208083 2013.

**Armstrong, M. and Y. Chen**, "Discount pricing," CEPR Discussion Paper 9327 2013.

**Autor, David et al.**, *Polanyi's paradox and the shape of employment growth*, National Bureau of Economic Research, 2014.

**Baliga, Sandeep and Jeffrey C Ely**, "Mnemonomics: the sunk cost fallacy as a memory kludge," *American Economic Journal: Microeconomics*, 2011, *3* (4), 35–67.

**Beauchamp, J.P., D.J. Benjamin, C.F. Chabris, and D.I. Laibson**, "How Malleable are Risk Preferences and Loss Aversion?," Harvard University Manuscript 2012.

**Beggs, Alan and Kathryn Graddy**, "Anchoring effects: Evidence from art auctions," *The American Economic Review*, 2009, pp. 1027–1039.

**Bénabou, Roland and Jean Tirole**, "Willpower and personal rules," *Journal of Political Economy*, 2004, *112* (4), 848–886.

humans is the aversion developed for foods which we consume prior to getting ill – an aversion that persists even if we know that the association is not causal (e.g., when the illness is due to chemotherapy).

**Bernheim, B Douglas and Antonio Rangel**, "Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics," *The Quarterly Journal of Economics*, 2009, *124* (1), 51–104.

**Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman**, "What's advertising content worth? Evidence from a consumer credit marketing field experiment," *The Quarterly Journal of Economics*, 2010, *125* (1), 263–306.

**Birnbaum, M.H.**, "Violations of monotonicity and contextual effects in choice-based certainty equivalents," *Psychological Science*, 1992, *3* (5), 310.

**_ , J.N. Patton, and M.K. Lott**, "Evidence against Rank-Dependent Utility Theories: Tests of Cumulative Independence, Interval Independence, Stochastic Dominance, and Transitivity," *Organizational Behavior and Human Decision Processes*, 1999, *77* (1), 44–83.

**Bodner, Ronit and Drazen Prelec**, "Self-signaling and diagnostic utility in everyday decision making," *The psychology of economic decisions*, 2003, *1*, 105–26.

**Brocas, I. and J.D. Carrillo**, "The brain as a hierarchical organization," *The American Economic Review*, 2008, *98* (4), 1312–1346.

**Brocas, Isabelle and Juan D Carrillo**, "Value computation and value modulation: a dual-process theory of self-control," Working Paper, USC 2013.

**Brown, Jennifer, Tanjim Hossain, and John Morgan**, "Shrouded attributes and information suppression: Evidence from the field," *The Quarterly Journal of Economics*, 2010, *125* (2), 859–876.

**Caplin, Andrew and Daniel Martin**, "A Testable Theory of Imperfect Perception," Working Paper 17163, National Bureau of Economic Research 2011.

**_ and _** , "Defaults and Attention: The Drop Out Effect," Working Paper 17988, National Bureau of Economic Research 2012.

**_ and Mark Dean**, "Rational inattention and state dependent stochastic choice," Working Paper 2013.

**Carbone, Enrica and John D Hey**, "A comparison of the estimates of EU and non-EU preference functionals using data from pairwise choice and complete ranking experiments," *Geneva Papers on Risk and Insurance Theory*, 1995, *20* (1), 111–133.

**Cavallo, V., M. Colomb, and J. Doré**, "Distance perception of vehicle rear lights in fog," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2001, *43* (3), 442–451.

**Chamley, C.P.**, *Rational herds: Economic models of social learning*, Cambridge University Press, 2003.

**Chetty, R., A. Looney, and K. Kroft**, "Salience and Taxation: Theory and Evidence," *American Economic Review*, 2009, *99* (4), 1145–1177.

**Chetty, Raj, Adam Looney, and Kory Kroft**, "Salience and Taxation: Theory and Evidence," Working Paper 13330, National Bureau of Economic Research 2007.

**Daw, Nathaniel D., Yael Niv, and Peter Dayan**, "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control," *Nature neuroscience*, 2005, *8* (12), 1704–1711.

**Erev, Ido and Ernan Haruvy**, "Learning and the economics of small decisions," *The handbook of experimental economics*, 2013, *2*.

**Evans, Jonathan St. B. T.**, "Dual-processing accounts of reasoning, judgment, and social cognition," *Annual Review of Psychology*, 2008, *59*, 255–278.

**Falk, Armin and Florian Zimmermann**, "A taste for consistency and survey response behavior," *CESifo Economic Studies*, 2013, *59* (1), 181–193.

**Feldman, Jacob**, "Bayesian Models of Perceptual Organization," in Johan Wagemans, ed., *Oxford Handbook of Perceptual Organization*, Oxford University Press, 2013.

**Fodor, Jerry A.**, *The Modularity of Mind: An Essay on Faculty Psycholgy*, The MIT Press, 1983.

**Fudenberg, D. and D.K. Levine**, "A dual-self model of impulse control," *The American Economic Review*, 2006, pp. 1449–1476.

**_ , _ , and Z. Maniadis**, "Reexamining Coherent Arbitrariness for the Evaluation of Common Goods and Simple Lotteries," Carlo F. Dondena Centre for Research on Social Dynamics 2010.

**Gabaix, X.**, "A sparsity-based model of bounded rationality," Technical Report, NYU Working Paper 2012.

**Gigerenzer, G. and W. Gaissmaier**, "Heuristic decision making," *Annual review of psychology*, 2011, *62*, 451–482.

**Gneezy, Uri, John A List, and George Wu**, "The uncertainty effect: When a risky prospect is valued less than its worst possible outcome," *The Quarterly Journal of Economics*, 2006, *121* (4), 1283–1309.

**Hey, John D**, "Does repetition improve consistency?," *Experimental economics*, 2001, *4* (1), 5–54.

**Hsee, C.K.**, "Less is better: when low-value options are valued more highly than high-value options," *Journal of Behavioral Decision Making*, 1998, *11* (2), 107–121.

**Iacobucci, Dawn**, *Kellogg on marketing*, John Wiley & Sons, 2001.

**Iyengar, S.S. and E. Kamenica**, "Choice proliferation, simplicity seeking, and asset allocation," *Journal of Public Economics*, 2010, *94* (7), 530–539.

**Jahedi, S.**, "A Taste for Bargains," Unpublished Working Paper 2011.

**Jones, M. and B.C. Love**, "Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition," *Behavioral and Brain Sciences*, 2011, *34* (04), 169–188.

**Kahneman, D.**, *Thinking, fast and slow*, Farrar, Straus and Giroux, 2011.

**_ and A. Tversky**, "Prospect theory: An analysis of decision under risk," *Econometrica: Journal of the Econometric Society*, 1979, pp. 263–291.

**_ and S. Frederick**, "A model of heuristic judgment," *The Cambridge handbook of thinking and reasoning*, 2005, pp. 267–294.

**Kamenica, E.**, "Contextual inference in markets: On the informational content of product lines," *American Economic Review*, 2008, *98*, 2127–2149.

**Keren, Gideon and Martijn C Willemsen**, "Decision anomalies, experimenter assumptions, and participants' comprehension: Revaluating the uncertainty effect," *Journal of Behavioral Decision Making*, 2009, *22* (3), 301–317.

**Larrick, Richard P**, "Debiasing," in D. J. Koehler and N. J. Harvey, eds., *Blackwell Handbook of Judgment and Decision-Making*, Blackwell Publishing, 2004, pp. 316–337.

**_** , "Broaden the decision frame to make effective decisions," *Handbook of principles of organizational behavior*, 2009, pp. 461–80.

**List, J.A.**, "Preference Reversals of a Different Kind: The 'More is Less' Phenomenon," *The American Economic Review*, 2002, *92* (5), 1636–1643.

**Loomes, Graham and Robert Sugden**, "Testing different stochastic specifications of risky choice," *Economica*, 1998, *65* (260), 581–598.

**Marr, David**, *Vision: A computational investigation into the human representation and processing of visual information*, Henry Holt and Co Inc., New York, NY, 1982.

**Mazar, N., B. Koszegi, and D. Ariely**, "True Context dependent Preferences? The Causes of Market dependent Valuations," *Journal of Behavioral Decision Making*, 2013, *27* (3), 200–208.

**McKenzie, Craig R. M. and Jonathan D. Nelson**, "What a Speaker's Choice of Frame Reveals: Reference Points, Frame Selection, and Framing Effects," *Psychonomic Bulletin and Review*, 2003, *10* (3), 596–602.

**Meer, Peter**, "Are we making real progress in computer vision today?," *Image and Vision Computing*, 2012, *30* (8), 472–473.

**Mueller-Frank, Manuel and Itai Arieli**, "Generic Outcomes of Observational Learning," SSRN Working Paper 2101661 2012.

**Mullainathan, S.**, "A memory-based model of bounded rationality," *The Quarterly Journal of Economics*, 2002, *117* (3), 735–774.

**Newell, Ben R. and David R. Shanks**, "Unconscious influences on decision making: a critical review," *Behavioral and Brain Sciences*, forthcoming.

**Peysakhovich, Alexander and D Rand**, "Habits of virtue: creating norms of cooperation and defection in the laboratory," SSRN Working Paper 2294242 2013.

**Polanyi, Michael**, *The Tacit Dimension*, Doubleday and Co., 1966.

**Predebon, J. et al.**, "Decrement of the Brentano Müller-Lyer illusion as a function of inspection time," *Perception*, 1998, *27*, 183–192.

**Prelec, D., B. Wernerfelt, and F. Zettelmeyer**, "The role of inference in context effects: Inferring what you want from what is available," *Journal of Consumer research*, 1997, *24* (1), 118–125.

**Pylyshyn, Zenon**, "Is vision continuous with cognition? The case for cognitive impenetrability of visual perception," *Behavioral and brain sciences*, 1999, *22* (3), 341–365.

**Pylyshyn, Zenon W**, *Computation and cognition*, Cambridge Univ Press, 1984.

**Rabin, Matthew and Georg Weizsacker**, "Narrow bracketing and dominated choices," *The American economic review*, 2009, *99* (4), 1508–1543.

**Rand, David G., Alexander Peysakhovich, Gordon T. Kraft-Todd, George E. Newman, Owen Wurzbacher, Martin A. Nowak, and Joshua D. Greene**, "Intuitive cooperation and the social heuristics hypothesis: evidence from 15 time constraint studies," *Available at SSRN: h ttp://ssrn. com/abstract*, 2013, *2222683.*

**Reber, A.S.**, "Implicit learning and tacit knowledge.," *Journal of experimental psychology: General*, 1989, *118* (3), 219.

**Ross, Helen**, "Mist, murk and visual perception," *New Scientist*, 1975, *66* (954).

**Rydval, Ondřej, Andreas Ortmann, Sasha Prokosheva, and Ralph Hertwig**, "How certain is the uncertainty effect?," *Experimental Economics*, 2009, *12* (4), 473–487.

**Schwartzstein, J.**, "Selective attention and learning," *Unpublished Manuscript, Dartmouth University*, 2012.

**Shaked, Moshe and J George Shanthikumar**, *Stochastic orders*, Springer, 2007.

**Shapiro, Jesse**, "A 'Memory-Jamming' Theory of Advertising," *Available at SSRN 903474*, 2006.

**Simonsohn, Uri**, "Direct risk aversion evidence from risky prospects valued below their worst outcome," *Psychological Science*, 2009, *20* (6), 686–692.

**Sims, C.A.**, "Rational inattention: a research agenda," Technical Report, Discussion paper Series 1/Volkswirtschaftliches Forschungszentrum der Deutschen Bundesbank 2005.

**Sinha, Pawan, Benjamin Balas, Yuri Ostrovsky, and Richard Russell**, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proceedings of the IEEE*, 2006, *94* (11), 1948–1962.

**Sloman, S.A.**, "The empirical case for two systems of reasoning.," *Psychological bulletin*, 1996, *119* (1), 3.

**Stanovich, K.E. and R.F. West**, "Individual differences in reasoning: Implications for the rationality debate?," *Behavioral and brain sciences*, 2000, *23* (5), 645–665.

**Starmer, Chris**, "Cycling with rules of thumb: An experimental test for a new form of non-transitive behaviour," *Theory and Decision*, 1999, *46* (2), 139–157.

**Tversky, A. and D. Kahneman**, "Judgment under uncertainty: Heuristics and biases," *Science*, 1974, *185* (4157), 1124–1131.

**_ and _**, "The framing of decisions and the psychology of choice," *Science*, 1981, *211* (4481), 453–458.

**_ and _**, "Rational choice and the framing of decisions," *The Journal of Business*, 1986, *59* (4), 251–278.

**Wernerfelt, B.**, "A rational reconstruction of the compromise effect: Using market data to infer utilities," *Journal of Consumer Research*, 1995, *21* (4), 627–633.

**Wilson, A.**, "Bounded memory and biases in information processing," *NAJ Economics*, 2002, *5* (3).

**Woodford, M.**, "Inattentive Valuation and Reference-Dependent Choice," *Unpublished Manuscript, Columbia University*, 2012.

# 6 Appendix: Learnability of $\alpha$ (For Online Publication)

The model in the text allows for the existence of knowledge held by System 1 which could *never* be discovered by System 2. Suppose there exist a pair $\alpha, \alpha' \in A$ such that, for every $x \in X$, $E[v|s(x), \alpha] = E[v|s(x), \alpha']$. Then System 2 could never discover whether $\alpha$ or $\alpha'$ holds, even though the distinction may be payoff-relevant, i.e. $\exists x \in X$, $E[v|x, \alpha] \neq E[v|x, \alpha']$.

In this section I note that if $\alpha$ is learnable by System 1 (in a particular sense) then $\alpha$ can be inferred by System 2 from System 1's responses. Therefore judgment will be unbiased when System 2 observes all of System 1's judgments, i.e. its judgments for every possible $s(x)$.

**Definition 1.** A distribution $f(v, x, \alpha)$ is *learnable* if $\forall \alpha, \alpha' \in A$, $\exists x \in X$,

$$E[v|s(x), \alpha] \neq E[v|s(x), \alpha'].$$

Learnability is a natural restriction if we think of System 1 as a naive learner: i.e., if System 1 simply stores the average observed $v$ for a given $s(x)$. Given an unlearnable distribution, there will always exist a coarsening of $A$ that is learnable (because, at worst, if $A$ is a singleton, then it is learnable).

The following proposition states that if System 2 can observe System 1's judgment for every element in $X$, and $f$ is learnable, then judgment will be unbiased.

**Proposition 4.** *If $f$ is learnable then for all $\alpha \in A$, $x \in X$, $m \in \mathbb{N}$, $\mathbf{x} \in X^m$, with $x' \in \mathbf{x} \iff x' \in X$,*

$$E_2^{\alpha,\mathbf{x}}(x) = E_P^{\alpha}(x).$$

*Proof.* Because $\mathbf{x}$ contains every element in $X$, then $\mathbf{E}_1$ will contain $E[v|s(x), \alpha]$ for every $x \in X$. Because $f$ is learnable, there is only a single $\alpha \in A$ that is consistent with this pattern, thus $E[\alpha|\mathbf{x}, \mathbf{E}_1] = \alpha$. Therefore $E_2 = E[v|\mathbf{x}, \mathbf{E}_1] = E[v|\mathbf{x}, \alpha] = E_P$. □

A related question is, how could System 2 have priors over $\alpha$, if $\alpha$ is a constant feature of the world? For example, suppose $v$ is the distance of an object, $x_i$ is the

blurriness of the object, and $\alpha_i$ is the correlation between blurriness and distance: then $\alpha_i$ is a function of physical constants and details about the atmosphere. Is it meaningful for there to be an objective distribution over possible values of $\alpha_i$? There is certainly no room for System 2 to *learn* the distribution $f(\alpha)$, since each universe has only one realization of $\alpha$. I think the answer to this can only be that $f(\alpha)$ represents System 2's priors, rather than an objective distribution, and that priors are an endowment. Similar questions apply to many applications of Bayesian reasoning, yet we seem to be able to talk sensibly about priors over fundamental parameters, such as gravitational constants, or the returns to education.

# 7 Appendix: Proofs & Additional Propositions (For Online Publication)

**Proof of Remark 1.**

*Proof.* We wish to show that System 2's bias is always weakly smaller, i.e.

$$Var[v - E[v|x, E[v|s(x), \alpha]]] \leq Var[v - E[v|s(x), \alpha]].$$

By the law of iterated expectations $E_2$ can be written as conditioning on a strictly larger information set than $E_1$.

$$E[v|s(x), \alpha] = E[v|s(x), E[v|s(x), \alpha]].$$

And it is easy to show that the variance of $(v - E[v|P])$ will always be smaller when $P$ constitutes a larger information set. □

**Proof of Remark 2.**

*Proof.* A bias exists if and only if there is some $x \in X$, $\alpha \in A$, such that $E[v|x, \alpha] \neq E[v|x, E[v|s(x), \alpha]]$. We can express $E_2$ as

$$\begin{aligned} E_2 &= E[v|x, E[v|s(x), \alpha]] \\ &= \int_{\bar{\alpha} \in A} E[v|x, \bar{\alpha}] f(d\bar{\alpha}|x, E[v|s(x), \alpha]). \end{aligned}$$

If the assumed condition holds, then for any $\bar{\alpha} \in A$ such that $f(\bar{\alpha}|x, E[v|s(x), \alpha]) > 0$ then $E[v|x, \bar{\alpha}] = E[v|x, \alpha]$, therefore:

$$\begin{aligned} &= \int_{\bar{\alpha} \in A} E[v|x, \alpha] f(d\bar{\alpha}|x, E[v|s(x), \alpha]) \\ &= E[v|x, \alpha] \\ &= E_P. \end{aligned}$$

Suppose the condition did not hold, then there would exist some $x \in X$, $\alpha, \alpha' \in A$

such that:

$$\begin{aligned} E[v|x,\alpha] &\neq E[v|x,\alpha'] \\ E[v|x,E[v|s(x),\alpha]] &= E[v|x,E[v|s(x),\alpha']]. \end{aligned}$$

Both cases give rise to the same $E_2$, but have different $E_P$, thus one case must be biased, i.e. either for $\alpha$ or $\alpha'$, $E_2 \neq E_P$. □

**Proof of Proposition 1.**

*Proof.* Suppose that there existed some $s \in S$ and $\alpha \in A$, such that there was a non-zero probability of bias, i.e.:

$$\int_{x\in X} 1\{E[v|x,\alpha] \neq E[v|x,E[v|s,\alpha]]\} f(x|s)dx \quad > 0.$$

This implies that either the probability of negative or the probability of positive bias is zero:

$$\int_{x\in \underline{X}} f(x|s)dx + \int_{x\in \bar{X}} f(x|s)dx > 0,$$

where $\underline{X} = \{x \in X : E[v|x,\alpha] < E[v|x,E[v|s,\alpha]]\}$ and $\bar{X} = \{x \in X : E[v|x,\alpha] > E[v|x,E[v|s,\alpha]]\}$. Now consider the element $\alpha'$ which is the highest ranked among those which have the same $E_1$, i.e. $\alpha' = \sup_{\succeq_s} \bar{A}$, where $\bar{A} = \{\bar{\alpha} \in A : E[v|s,\bar{\alpha}] = E[v|s,\alpha]\}$. Then by definition, for all $x \in X$ with $s(x) = s$,

$$\begin{aligned} E[v|x,\alpha'] &\geq \int_{\bar{\alpha}\in\bar{A}} E[v|x,\bar{\alpha}] f(\bar{\alpha}|x,E[v|s(x),\alpha])d\bar{\alpha} \\ &= E[v|x,E[v|s(x),\alpha]], \end{aligned}$$

so for $\underline{X}$,

$$\forall x \in \underline{X},\ E[v|x,\alpha] < E[v|x,\alpha'],$$

meaning that if $\int_{x\in\underline{X}} f(x|s)dx > 0$ then

$$\int_{x\in\underline{X}} \left(E[v|x,\alpha'] - E[v|x,\alpha]\right) f(x|s)dx \quad > \quad 0,$$

and therefore

$$\int_{x\in X} \left(E[v|x,\alpha'] - E[v|x,\alpha]\right) f(x|s)dx > 0.$$

However this contradicts our assumption that $\alpha$ and $\alpha'$ both yield the same $E_1$, i.e. that:

$$\int_{x\in X} E[v|x,\alpha']f(x|s)dx = \int_{x\in X} E[v|x,\alpha]f(x|s)dx.$$

A symmetric argument will also deliver a contradiction if $\int_{x\in \bar{X}} f(x|s)dx > 0$ (using the infinum of $\bar{A}$ instead of its supremum). Thus for any given $s$ and $\alpha$ the probability of bias must be zero, therefore the unconditional probability of bias must be zero. □

**Unambiguous $f$ when $\alpha$ is separable.**

If $\alpha$ is a vector of real numbers, with log concave prior distributions, and $E[v|x,\alpha]$ is linear in each $\alpha_i$, then the MLRP will hold between $E_1$ and each $\alpha_i$. This implies that an increase in $E_1$ will cause System 2 to increase their posteriors over every $\alpha_i$ (in the sense of stochastic dominance), which in turn implies that $f$ is unambiguous.

**Proposition 5.** *If $A = \mathbb{R}^n$, $\alpha$ and $x$ are independent, and*

$$E_1 = E[v|s(x),\alpha] = \sum_{i=1}^{n} \alpha_i g(s(x)),$$

*and each $\alpha_i$ is distributed independently with $F(\alpha_i|s(x))$ differentiable and $f(\alpha_i|s(x))$ log-concave, and $E[v|x,\alpha]$ is increasing in each $\alpha_i$, then $f$ is unambiguous.*

*Proof.* Shaked and Shanthikumar (2007) Theorem 6.B.9 establishes that the posteriors over $\alpha_i$ will increase in $E_1$, in the sense of stochastic dominance. Because $E_P$ is increasing in each $\alpha$, then $E_2$ must increase, thus $f$ is unambiguous. □

**Example 1. (Example of $f$ with bias)** Let $v, \alpha, x \in \{0,1\}$, with $f(v=1) = \frac{1}{2}$, and $s(x) = 0$. Suppose that if $v = 0$ then $\alpha$ and $x$ are uniformly distributed and independent, but if $v = 1$ then with equal probability $\alpha = x = 1$, or $\alpha = x = 0$. I.e., for all $\alpha, x \in \{0,1\}$:

$$f(\alpha, x|v=0) = \tfrac{1}{4}, \qquad f(\alpha, x|v=1) = \begin{cases} \frac{1}{2} & , \alpha = x, \\ 0 & , \alpha \neq x. \end{cases}$$

Then we can write:

$$\begin{aligned} E_P = E[v|\alpha, x] &= \begin{cases} \frac{2}{3} & , \alpha = x \\ 0 & , \alpha \neq x \end{cases} \\ &= \frac{2}{3}(1 - \alpha - x + 2\alpha x) \\ E_1 = E[v|\alpha] &= \sum_{x=0}^{1} E[v|\alpha, x] f(x|\alpha) \\ &= 0 \times \frac{1}{4} + \frac{2}{3} \times \frac{3}{4} = \frac{1}{2} \\ E_2 = E[v|E[v|\alpha], x] &= \frac{1}{2}. \end{aligned}$$

Here the pooled-information expectation includes an interaction term between $\alpha$ and $x$. In this case we do not know whether a realization of $\alpha$ represents good news or bad news about $v$ until we know the realization of $x$. In fact, in this case it means that the intermediate expectation, $E_1$, will be entirely uninformative, because $E_1 = \frac{1}{2}$ everywhere, independent of $\alpha$. System 2 cannot learn anything about $\alpha$, so both System 1 and 2 will be biased relative to the pooled-information benchmark (i.e., $\forall \alpha \in A, x \in X$, $E_1 = E_2 \neq E_P$). □

**Proof of Proposition 2.**

*Proof.* The result will follow if we can show that System 2 interprets $E_1$ in the same way, for both $x$ and $x'$, i.e. if, for a given $E_1$:

$$E[v|x, E_1] = E[v|x', E_1],$$

This will follow because $E_2$ can be written as:

$$\begin{aligned} E[v|x, E_1] &= \int \bar{E}_P f(\bar{E}_P|x, E_1) d\bar{E}_P \\ &= \int \bar{E}_P \frac{f(\bar{E}_P, E_1|x)}{f(E_1|x)} d\bar{E}_P \\ &= \int \bar{E}_P \frac{f(\bar{E}_P, E_1|x')}{f(E_1|x')} d\bar{E}_P \\ &= E[v|x', E_1]. \end{aligned}$$

Because $f$ is assumed to be congruent an increase in $E_1$ must also weakly increase $E_2$, thus the conclusion is direct. □

**Proof of Corollary 2.**

*Proof.* When $m = 1$ then $\bar{X} = \begin{pmatrix} \bar{x}_i \\ \dots \\ \bar{x}_n \end{pmatrix}$, and $\hat{X} = \begin{pmatrix} \hat{x}_i \\ \dots \\ \hat{x}_n \end{pmatrix}$ then:

$$\begin{aligned} E_P &= \sum_{i=1}^{n} \bar{x}_i \hat{x}_i \alpha_i \\ E_1 &= \sum_{i=1}^{n} \bar{x}_i \alpha_i, \end{aligned}$$

and

$$\begin{aligned} (\bar{X}\Omega\bar{X}') &= \sum_{i=1}^{n} \bar{x}_i^2 \sigma_i^2 \\ E[\alpha|E_1]_i &= E[\alpha_i] + \frac{\bar{x}_i \sigma_i^2}{\sum_{j=1}^{n} \bar{x}_j^2 \sigma_j^2} \left( E_1 - \sum_{j=1}^{n} \bar{x}_j E[\alpha_j] \right) \\ E_2 &= \sum \bar{x}_i \hat{x}_i E[\alpha_i|E_1] \\ &= \sum_{i=1}^{n} \bar{x}_i \hat{x}_i E[\alpha_i] + \sum_{i=1}^{n} \bar{x}_i \hat{x}_i \frac{\bar{x}_i \sigma_i^2}{\sum_{j=1}^{n} \bar{x}_j^2 \sigma_j^2} \left( E_1 - \sum_{j=1}^{n} \bar{x}_j E[\alpha_j] \right) \\ &= \sum_{i=1}^{n} \bar{x}_i \hat{x}_i E[\alpha_i] + \left( \sum_{i=1}^{n} \frac{\bar{x}_i^2 \sigma_i^2}{\sum_{j=1}^{n} \bar{x}_j^2 \sigma_j^2} \hat{x}_i \right) \left( \sum_{k}^{n} \bar{x}_k (\alpha_k - E[\alpha_k]) \right) \\ E_2 - E_P &= \sum_{k}^{n} \bar{x}_k (\alpha_k - E[\alpha_k]) \left( \sum_{i=1}^{n} \frac{\bar{x}_i^2 \sigma_i^2}{\sum_{j=1}^{n} \bar{x}_j^2 \sigma_j^2} \hat{x}_i - \hat{x}_k \right). \end{aligned}$$

□

**Proof of Corollary 3.**

*Proof.*

$$\begin{aligned}
\bar{X}\Omega\bar{X}' &= \begin{pmatrix} \Sigma_{\bar{x}^2\sigma^2} & \Sigma_{\bar{x}\bar{x}'\sigma^2} \\ \Sigma_{\bar{x}\bar{x}'\sigma^2} & \Sigma_{\bar{x}'^2\sigma^2} \end{pmatrix} \\
(\bar{X}\Omega\bar{X}')^{-1} &= \frac{1}{\Sigma_{\bar{x}^2\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - (\Sigma_{\bar{x}\bar{x}'\sigma^2})^2} \begin{pmatrix} \Sigma_{\bar{x}'^2\sigma^2} & -\Sigma_{\bar{x}\bar{x}'\sigma^2} \\ -\Sigma_{\bar{x}\bar{x}'\sigma^2} & \Sigma_{\bar{x}^2\sigma^2} \end{pmatrix} \\
(\bar{X}\circ\hat{X})\Omega\bar{X}' &= \begin{pmatrix} \Sigma_{\bar{x}^2\hat{x}\sigma^2} & \Sigma_{\bar{x}\bar{x}'z\sigma^2} \\ \Sigma_{\bar{x}\bar{x}'\hat{x}'\sigma^2} & \Sigma_{\bar{x}'^2\hat{x}'\sigma^2} \end{pmatrix} \\
\bar{X}(\alpha - E[\alpha]) &= \begin{pmatrix} \Sigma_{i=1}^n \bar{x}_{1,i}(\alpha_i - E[\alpha_i]) \\ \Sigma_{i=1}^n \bar{x}_{2,i}(\alpha_i - E[\alpha_i]) \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
&(\bar{X}\circ\hat{X})\Omega\bar{X}'(\bar{X}\Omega\bar{X}')^{-1}\bar{X}(\alpha - E[\alpha]) = \\
= \; & \frac{1}{D}\begin{pmatrix} \Sigma_{\bar{x}^2\hat{x}\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - \Sigma_{\bar{x}\bar{x}'\hat{x}\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2} & -\Sigma_{\bar{x}^2\hat{x}\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2} + \Sigma_{\bar{x}\bar{x}'\hat{x}\sigma^2}\Sigma_{\bar{x}^2\sigma^2} \\ \Sigma_{\bar{x}\bar{x}'\hat{x}'\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - \Sigma_{\bar{x}'^2\hat{x}'\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2} & -\Sigma_{\bar{x}\bar{x}'\hat{x}'\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2} + \Sigma_{\bar{x}'^2\hat{x}\sigma^2}\Sigma_{\bar{x}^2\sigma^2} \end{pmatrix}\begin{pmatrix} \Sigma_{\bar{x}(\alpha-E[\alpha])} \\ \Sigma_{\bar{x}'(\alpha-E[\alpha])} \end{pmatrix} \\
= \; & \frac{1}{D}\begin{pmatrix} \left(\Sigma_{\bar{x}^2\hat{x}\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - \Sigma_{\bar{x}\bar{x}'\hat{x}\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2}\right)\Sigma_{\bar{x}(\alpha-E[\alpha])} + \left(-\Sigma_{\bar{x}^2\hat{x}\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2} + \Sigma_{\bar{x}\bar{x}'\hat{x}\sigma^2}\Sigma_{\bar{x}^2\sigma^2}\right)\Sigma_{\bar{x}'(\alpha-E[\alpha])} \\ \left(\Sigma_{\bar{x}\bar{x}'\hat{x}'\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - \Sigma_{\bar{x}'^2\hat{x}'\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2}\right)\Sigma_{\bar{x}(\alpha-E[\alpha])} + \left(-\Sigma_{\bar{x}\bar{x}'\hat{x}'\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2} + \Sigma_{\bar{x}'^2\hat{x}\sigma^2}\Sigma_{\bar{x}^2\sigma^2}\right)\Sigma_{\bar{x}'(\alpha-E[\alpha])} \end{pmatrix}
\end{aligned}$$

where

$$D = \Sigma_{\bar{x}^2\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - (\Sigma_{\bar{x}\bar{x}'\sigma^2})^2,$$

and

$$(\bar{X}\circ\hat{X})(\alpha - E[\alpha]) = \begin{pmatrix} \Sigma_{\bar{x}\hat{x}(\alpha-E[\alpha])} \\ \Sigma_{\bar{x}'\hat{x}'(\alpha-E[\alpha])} \end{pmatrix},$$

hence solving for $E_2 - E_P$ for the case $(\bar{x}, \hat{x})$ gives us:

$$\begin{aligned}
&E_2 - E_P = \\
&\frac{\Sigma_{\bar{x}^2\hat{x}\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - \Sigma_{\bar{x}\bar{x}'\hat{x}\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2}}{\Sigma_{\bar{x}^2\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - (\Sigma_{\bar{x}\bar{x}'\sigma^2})^2}\Sigma_{\bar{x}(\alpha-E[\alpha])} - \frac{\Sigma_{\bar{x}^2\hat{x}\sigma^2}\Sigma_{\bar{x}\bar{x}'\sigma^2} - \Sigma_{\bar{x}\bar{x}'\hat{x}\sigma^2}\Sigma_{\bar{x}^2\sigma^2}}{\Sigma_{\bar{x}^2\sigma^2}\Sigma_{\bar{x}'^2\sigma^2} - (\Sigma_{\bar{x}\bar{x}'\sigma^2})^2}\Sigma_{\bar{x}'(\alpha-E[\alpha])} - \Sigma_{\bar{x}\hat{x}(\alpha-E[\alpha])}.
\end{aligned}$$

□