Cardinal Representations of Information*

Jeffrey Mensch†

November 20, 2017

PRELIMINARY

Abstract

In the spirit of von Neumann and Morgenstern (1947), this paper provides an axiomatic representation of information. Under the von Neumann-Morgenstern axioms, along with an additional continuity, indifference to randomization, and a Blackwell informativeness axiom, I show that any ordering over information can be essentially uniquely represented as, equivalently: (a) a strictly increasing cost of information acquisition; (b) for a given prior, the expected utility from a decision problem; (c) for a given prior, an additive posterior-separable measure of uncertainty; and (d) a separable cost of signals. I discuss the implications of the results for the rational inattention literature.

*I would like to thank Robert Barsky, Alessandro Pavan, Marciano Siniscalchi, Bruno Strulovici, Yuta Takahashi, and Asher Wolinsky for helpful conversations regarding both the general approach to information, and how it relates to the rational inattention literature. All remaining errors are my own.

†Hebrew University; jeffrey.mensch@mail.huji.ac.il

1 Introduction

Information is fundamental to the study of economics, as it determines the incentives that agents have to make choices. Since Blackwell (1951, 1953), economists have defined information via experiments about an unknown, underlying state of nature. In the absence of an experiment, one has a garbled view of the underlying state; the experiment removes the garbling. However, up to now, there has not been a formal study of the representation of the full set of possible information structures; the Blackwell order is merely a partial order. For a decision maker (DM) who is deciding which experiment to run to acquire information, it is necessary to be able to compare *all possible* experiments, or at the very least, experiments that are not Blackwell-ordered. While such experiments have appeared in other models, their representations have not been derived from a primitive, axiomatic framework, without built-in representational assumptions. This paper serves to provide a unified axiomatic basis for these frameworks, helping to shed additional light on how to properly model information.

A key feature of experiments (in the Blackwell sense) is that they are defined with respect to a state space and a signal space, but without reference to a particular prior distribution. Thus, one experiment will be more informative than another regardless of the prior over states. If one is to properly model information acquisition, then, one should respect this prior-independence property. The representations here, therefore, will maintain a consistent ranking of the informativeness of experiments for any prior.

There are several representations that one can consider, each of which have appealing features. The most basic one is a simple binary relation between possible experiments. Naturally, one would want it to be consistent with the Blackwell order. If one wants a cardinal representation, then it should also satisfy conditions akin to the von Neumann-Morgenstern axioms.

Another possible representation, consistent with the spirit of Blackwell (1953), is that information is relevant for a particular decision problem.[1] Thus, one could compare experiments by how much a decision maker would value them. This will, then, serve as a completion of the Blackwell partial order.

Information may also be costly to acquire. Therefore, there may be an additional price that a DM might have to pay to run a particular experiment. I consider three different representations of such costs, including that found in the rational attention literature of measuring attention costs by changes in posterior-separable measures of uncertainty.

The main result of the paper is to show that all of these representations are equivalent

[1]Note that in Blackwell (1953), an experiment is more informative if it leads to high expected utility in *all* decision problems, rather than a particular one.

and essentially unique, under the additional axiom of "indifference to randomization" (IR). Roughly, this axiom states that if an experiment π^* generates the same posteriors as from randomizing over some set of experiments $\{\pi\}$, then the two ways of generating these posteriors are equivalent. That this yields a posterior-separable representation is somewhat surprising, since it is not obviously equivalent to, say, an independence axiom, which is often used to generate separable representations. I use a constructive proof to derive the representation, which allows for easy translation from a general cost function for a specific experiment π to its representation in terms of posteriors μ.

The representations in this paper allow for discussion of comparative statics over preferences for information. I show that a higher cost of information acquisition is equivalent to having a more convex measure of uncertainty. Since the measure of uncertainty can also be seen as the expected utility from a decision problem, one can think of this as saying that one faces additional risk from one's decision. This allows for a better understanding of the connection between information and risk.

While the representability of information by a posterior-separable measure of uncertainty may seem to lend credence to the use of entropy reduction as a proxy for costly information acquisition, such a conclusion needs to be taken with a grain of salt. This is because this representation of information depends on which prior one uses. As I show in Section 5 through a principal-agent problem, a naive use of the rational inattention approach, which does not take this prior dependence into account, can lead to unacceptable conclusions: it becomes costless to monitor an agent's effort. By contrast, when used properly, the methods derived in the main representation theorem yield a tractable method of solving the problem, while properly incorporating the tradeoffs of costly monitoring.

Lastly, I discuss some of the hazards associated with associating entropy reduction with costly information acquisition. I argue that there is some confusion in the literature regarding the relationship between information as defined by Shannon (1948) and as defined by Blackwell (1951, 1953). While the former treats both the state and the signal symmetrically, Blackwell experiments treat them asymmetrically, in that there is some technology which generates the signal from the state. I argue that the latter is more appropriate when modeling information acquisition.

1.1 Related literature

This paper takes an axiomatic decision-theoretic approach to representing information. As such, it is most closely related to other papers that use the same methodology. The seminal work of von Neumann and Morgenstern (1947) has provided the axioms needed for a

cardinal representation. Other works, which use similar techniques that consider preferences over menus, include Kreps (1979) and Dekel et al. (2001). While this paper does not consider menus, the techniques here are similar, and the (IR) axiom here has an analogue in the latter.

This literature has led others to consider decision makers who are constrained by additional costs of optimization, of which information acquisition is an example. Ergin and Sarver (2010) provide axioms for a DM who has a cost of considering additional options in a set. They show that there is a cardinal representation with an additively separable contemplation cost whenever the DM exhibits "aversion to contingent planning" over menus (along with some other technical axioms).

Perhaps the paper that is most closely related to this one is that of De Oliveira et al. (2016) who build on the framework of Ergin and Sarver to explicitly provide an axiomatic foundation for a DM with costly information acquisition. They use preferences over menus to identify the implicit costs of experiments. Lastly, they provide a comparative statics result, showing that DM 1 acquires more information than DM 2 if and only if DM 1 prefers a lottery over menus with a higher information premium whenever DM 2 does. Their paper is distinct from this one in several ways, stemming from the fact that the primary object of consideration in their axiomatic framework is menus over acts/choices, whereas that here is experiments. Thus, their cardinal representation result is that of preferences over acts with an implicit, hidden information cost. Here, the main result is a cardinal representation of information, which need not be connected to a particular decision problem. However, if desired, it can be interpreted as a cost of information acquisition for an implicit, hidden decision problem, or as the expected utility for such a decision problem. This also yields the distinction in our comparative statics results: while they look at comparisons of preferences over lotteries over *menus* to derive comparative statics on information costs, I look at comparisons of lotteries over *experiments* to derive comparative statics on the shape of the implicit utility functions or information costs.

The ability to generate arbitrary informative experiments about an underlying state has been incorporated into several economic frameworks. For instance, in the Bayesian persuasion literature (Kamenica and Gentzkow, 2011), there is a sender who can commit ex-ante to generate any arbitrary information structure in order to convince a receiver to take a particular action conditional on the state. They extend this to situations where it is costly to reveal more information, using a posterior-separable measure of uncertainty as their cost function (Kamenica and Gentzkow, 2014).

Much of the motivation for this paper is from the rational inattention literature, so I highlight a few key papers here. Sims (1998, 2003) suggests that the presence of sticky prices may be due to agents being unable to process all information that is available. He models this

by allowing agents to reduce Gaussian noise through costly information acquisition, which he models by reduction of informational entropy/mutual information as defined by Shannon (1948) and Kullback and Leibler (1951). Mackowiak and Wiederholt (2009) build on this, introducing tradeoffs in attention allocation as well as feedback effects, and show how their model reacts to both private and aggregate shocks. They then calibrate their model to microeconomic data, and show that their model can predict many of the observed effects.

There has also been a recent literature on flexible information acquisition by a rationally inattentive agent. Woodford (2008) considers a binary choice model in which players can acquire arbitrary signals, whose costs are measured by entropy reduction. Matejka and McKay (2015) extend this to arbitrary finite action spaces, showing that rational inattention can provide a basis for the multinomial logit choice model. Caplin and Dean (2013) extend this approach to any posterior-separable measure of uncertainty, while in their (2015) paper, they examine such decision problems from a revealed preference approach. Other papers that use this approach include Cabrales et al. (2013), Ravid (2017), Denti (2017), and Mensch (2017).

A recent, closely related paper by Caplin, Dean, and Leahy (2017) provides an axiomatic foundation for the representation of information acquisition costs by a posterior-separable (PS) measure of uncertainty. The results here would be considered a PS representation, but not what they call a "uniform posterior-separable" (UPS) representation, in which the measure of uncertainty does not depend on the prior (which I will show is inconsistent with a representation of Blackwell experiments). Nevertheless, the prior-independence of the cost in this paper provides additional structure beyond being merely posterior-separable when comparing the representations across priors. An additional important difference is that, while I consider Blackwell experiments as the primitives for comparison, Caplin et al. consider state-dependent stochastic choice (SDSC) data. Given the different foundations, the two papers should be considered complementary.

2 Representations of Information

Let Ω be a finite state space of size N. An *experiment* is defined as a mapping $\pi : \Omega \to \Delta(\mathcal{S})$, where S is a signal space. Let the set of all possible experiments be Π. I formally define the topology over Π in Appendix A. For a given prior $\mu_0 \in \text{int}(\Delta(\Omega))$, one would then use Bayes' rule to update one's beliefs: that is, for a signal s that occurs with positive probability given π, one would have

$$\tau_\pi(s)\mu(\omega|s) = \mu_0(\omega)\pi(s|\omega)$$

where

$$\tau_\pi(s) = \sum_{\omega \in \Omega} \mu_0(\omega)\pi(s|\omega)$$

While, in general, $\mathcal{S}$ could be any compact metric space, it will become clear that it will be without loss of generality for our purposes to assume that $\mathcal{S} = \Delta(\Omega)$, with s, for a given prior μ_0, being the induced posterior: $\mu(\cdot|s) = s$. Following Blackwell (1951), we call such experiments "standard experiments." This also allows one to write $\tau_\pi(\mu(\cdot|s)) \equiv \tau_\pi(s)$, and so $\tau_\pi \in \Delta(\Delta(\Omega))$. It then follows that the representation of an experiment by π or, for a given prior, by τ_π, is equivalent. I formalize this in the following lemma.[2]

Lemma 1: *Within the class of standard experiments, given τ_π and prior μ_0, there exists an essentially unique π which generates τ_π, i.e. if some other π' generates τ_π, then π and π' must disagree on a set of measure* 0.

Proof: See Appendix A.

There are many ways in which economists have thought about representing information. In each of the following subsections, I formally define each of these representations.

2.1 Binary Relation

The most basic way to represent information is through a binary relation $\succsim$ over the full set of possible experiments, which captures the notion of "more information" or "less information." Of course, in order to coherently define information, this ordering must satisfy certain axioms. I list these below.

The first three axioms are essentially the standard von Neumann-Morgenstern axioms of expected utility theory.

Axiom 1: Completeness and Transitivity

(a) For all π_1, π_2, either $\pi_1 \succsim \pi_2$ or $\pi_2 \succsim \pi_1$.

(b) If $\pi_1 \succsim \pi_2$ and $\pi_2 \succsim \pi_3$, then $\pi_1 \succsim \pi_3$.

Axiom 2: Independence

If $\pi_1 \succsim \pi_2$, then for any π' and $\alpha \in (0, 1)$,

$$\alpha\pi_1 + (1-\alpha)\pi' \succsim \alpha\pi_2 + (1-\alpha)\pi'$$

Axiom 3: Archimedean Continuity

[2]Kamenica and Gentzkow (2011) show a similar result, but only do so for finite signal realizations, since it is without loss of generality in their environment to consider information that is useful for a finite decision problem. Since I abstract away from any single decision problem, one must extend this to infinite signals.

For any π_1, π_2, π_3 such that $\pi_1 \succ \pi_2 \succ \pi_3$, there exists $\alpha \in (0,1)$ such that $\pi_2 \sim \alpha\pi_1 + (1-\alpha)\pi_3$.

Axiom 3 is the standard continuity concept from the von Neumann-Morgenstern axioms, and is frequently just referred to as "continuity." However, when considering information, this continuity concept will not suffice, since this tells us how to compare lotteries over experiments, but does not allow comparison of experiments that lead to similar outcomes. Since there are a potentially uncountable number of possible experiments, such regularity is necessary if we are to have well-behaved cardinal representations of information. I therefore introduce Axiom 4, which ensures that experiments are well-behaved in this way as well.

Axiom 4: Belief Continuity

Fix prior $\mu_0 \in int(\Delta(\Omega))$, and experiment π. Then the upper- and lower-contour sets of $\succsim$ over π are closed in the weak topology as defined over $\tau_{\pi'} \in \Delta(\Delta(\Omega))$, i.e. the induced distribution of posteriors given experiment π'.*

One potential concern with the definition of continuity in Axiom 4 is that it appears to be prior dependent: continuity is defined in reference to some prior μ_0. Thus one may be worried that the upper- and lower-contour sets may be closed with respect to some prior μ_0, but not some other prior μ_0'. To alleviate concerns that this is a substantive assumption, I show in the following lemma that if continuity holds with respect to one prior, then it must hold with respect to all.

Lemma 2: *For a given experiment π, if the upper- and lower-contour sets of $\succsim$ are closed given $\mu_0 \in int(\Delta(\Omega))$, then they are closed for any $\mu_0' \in int(\Delta(\Omega))$.*

Proof: See Appendix A.

It is now clear why one can restrict $\mathcal{S} = \Delta(\Omega)$: what matters for assessing the ordering over experiments is the distribution of beliefs they induce. Thus, if some signal s generates belief μ, it is without loss of generality to just label s as μ.

The next axiom provides an important criterion for information: strictly more informative signals should be ranked strictly higher. The standard definition of information used in economics is that of Blackwell (1951, 1953), presented below.

Definition 1: A signal π is *Blackwell (strictly) more informative* than signal π', if given prior μ_0, for every (strictly) convex continuous function $\phi : \Delta(\Omega) \to \mathbb{R}$,

$$\int_{\Delta(\Omega)} \phi(\mu) d\tau_\pi(\mu) \geq (>) \int_{\Delta(\Omega)} \phi(\mu) d\tau_{\pi'}(\mu)$$

Axiom 5: Blackwell monotonicity

If π is strictly more informative than π', then $\pi \succ \pi'$.

It is well-known that the Blackwell order is independent of the prior $\mu_0 \in \text{int}\Delta(\Omega)$. To see this, one can restrict one's attention to standard experiments, in which case Definition 1 is equivalent to saying that given π, π', there exists some experiment $\tilde{\pi} : \mathcal{S} \to \Delta(\mathcal{S}')$ such that

$$\pi'(s'|\omega) = \sum_{s \in \mathcal{S}} \tilde{\pi}(s'|s)\pi(s|\omega)$$

Belief continuity and Blackwell monotonicity do not, however, rule out the possibility that there could be a set of experiment $\{\pi\}$ such that randomizing over them yields the same distribution over beliefs as some π', which would then be ranked differently. To rule this out, the next axiom states that if one experiment is Blackwell equivalent (in terms of induced beliefs) to randomizing over other experiments, then the former must be ranked the same as the latter randomization. The basic idea justifying this is that information is useful to a decision maker (DM) insofar as the DM can make an optimal choice contingent on the information that he has. Hence one can always randomize over one's choice of information acquisition in order to generate a particular distribution of beliefs, and how one gets there does not affect its usefulness.

Axiom 6: Indifference to Randomization (IR)

Suppose that, for some probability measure λ over Π, one has, for some π^*,

$$\tau_{\pi^*}(\mu) = \int_{\Pi} \tau_{\pi}(\mu) d\lambda(\pi)$$

Then $\pi^* \sim \int_{\Pi} \pi d\lambda(\pi)$.

One can therefore view the characterization here as a completion of the Blackwell order over experiments. Of course, Axioms 1-6 are not the only possible way to represent information. In the following subsections, I define several other possible ways.

2.2 Cost of information acquisition

One can also think about information in the following sense. A DM can acquire information flexibly, in that any informative experiment is possible. However, the more informative the experiment, the more costly it is to run. Formally, there is some function

$$c : \Pi \to \mathbb{R}$$

which is continuous and strictly increasing in the Blackwell order. I assume that the order induced by c satisfies (IR), as defined in the previous section, in the same spirit that any

lottery over experiments that provides more information, even probabilistically, should cost more. Lastly, any uninformative experiment is costless.

If any $\tilde{c}$ which induces the same order as c is a positive linear transformation of c, we say that c is *essentially unique*, and that $\tilde{c}$ is *essentially equivalent* to c.

2.3 Utility of decision problem

Another way of modeling information is by how beneficial it is to a DM. Let X be the set of actions that the DM can choose from, and $u : X \times \Omega \to \mathbb{R}$ define the ex-post utility from the choice of action $x \in X$ given state ω. As has been well-known since Blackwell (1953), a DM prefers π over π' for all decision problems u, given μ_0,

$$U(\pi) \equiv E_{<\mu|\pi>}[\max_{x \in X} E_\mu[u(x,\omega)]|\mu_0] \geq E_{<\mu|\pi'>}[\max_{x \in X} E_\mu[u(x,\omega)]|\mu_0] \equiv U(\pi') \tag{1}$$

if and only if π is Blackwell more informative than π'.

The Blackwell order, however, is merely a partial order. Since we are interested in comparing *all* experiments, Blackwell's results do not suffice.

One potential way to induce a complete order over experiments is to compare experiments for a given decision problem u, starting from prior μ_0. That is, if (1) holds for a given u, then π is ranked higher than π'. To make this consistent with the Blackwell order, I require that if π is strictly Blackwell more informative than π', then the above inequality is strict.

One potential question to ask here is whether the ordering defined by a given u is unique to that utility function. Clearly, in a strict sense, the answer is no; any affine transformation of u will induce the same order. Thus the question should be phrased as whether u is *essentially unique*. I introduce three criteria that define essential uniqueness.

1. Affine transformations of u are irrelevant, so if if $\tilde{u} = au + b$ for some positive a, then $\tilde{u}$ is essentially equivalent to u.

2. Additional actions that are never optimal are irrelevant, i.e. if there is some $\tilde{u} : \tilde{X} \times \Omega \to \mathbb{R}$ such that $X \subset \tilde{X}$, $\tilde{u}(x,\omega) = u(x,\omega)$ for $x \in X$, and for any $\mu \in \Delta(\Omega)$, $\exists x \in X$ s.t. $x \in \arg\max_{\tilde{x}} \sum_{\omega \in \Omega} \tilde{u}(\tilde{x},\omega)\mu(\omega)$, then $\tilde{u}$ is essentially equivalent to u.

3. Relabeling the action space is irrelevant: if $\tilde{u} : \tilde{X} \times \Omega \to \mathbb{R}$ such that there exists a 1-to-1 mapping $\varphi : \tilde{X} \to X$ such that $\tilde{u}(\tilde{x},\omega) = u(x,\omega)$ for all ω, then $\tilde{u}$ is essentially equivalent to u.

If any utility function $\tilde{u}$ that induces the same ordering over Π as u can be reduced to u through these three criteria, then we say that u is *essentially unique*.

2.4 Posterior-separable measure of uncertainty

In recent years, a popular way of representing information has been through a *posterior-separable measure of uncertainty*. This first appeared in the literature on rational inattention, where Sims (2003) suggested measuring the cost of information acquisition by

$$c(\pi) = E_{<\mu|\pi>}[H(\mu)] - H(\mu_0) \tag{2}$$

where $H(\mu) = -\sum_{\omega\in\Omega} \mu(\omega) \ln \mu(\omega)$, i.e. informational entropy as defined by Shannon (1948). More recently, others have noted that (2) will give a cost function that is strictly increasing in the Blackwell order for any strictly convex function H, which has been called a "measure of uncertainty" (Caplin and Dean, 2013; Kamenica and Gentzkow, 2014; Ely et al., 2015).

Part of the appeal of defining the cost in this form is that it is *posterior-separable*: that is, one can look at the realization of H, posterior-by-posterior, rather than looking at the entire experiment π. This allows for easy comparison of experiments for a DM who is acquiring information flexibly; the DM can assess the tradeoffs between any two given posteriors $\mu, \mu' \in \text{supp}(\tau)$ by examining the effect of a perturbation on H, without needing to worry about effects from the other $\tilde{\mu} \in \text{supp}(\tau) \setminus \{\mu, \mu'\}$. Thus one can use a straightforward first-order approach to solve the DM's decision problem as to how acquire information (Caplin and Dean, 2013; Matejka and McKay, 2015; Yang, 2015; Mensch, 2017).

As with the interpretation of information as a cost of information acquisition in Subsection 1.2, one can ask if this representation is unique. I define H to be *essentially unique* if any $\tilde{H}$ that induces the same ordering over experiments can be written as

$$\tilde{H}(\mu) = \lambda_0 H(\mu) + \sum_{n=1}^{N} \lambda_n \mu(\omega_n)$$

for $\lambda_0 \in \mathbb{R}_{++}$ and $\lambda_n \in \mathbb{R}$. This is to ensure that the cost of a given signal (as given by the expected difference in H) remains the same in all representations.

2.5 Cost of signals

Recall that by Bayes' rule, for any signal s that occurs with positive probability,

$$\tau_\pi(s)\mu(\omega|s) = \mu_0(\omega)\pi(s|\omega)$$

As long as $\frac{\pi(s|\omega)}{\tau_\pi(s)}$ remains the same for two different experiments, they induce the same posterior μ given μ_0. In particular, this implies that all that is needed for π' to generate μ is that $\pi'(s|\omega)/\pi'(s|\omega') = \pi(s|\omega)/\pi(s|\omega')$ for all ω, ω'. As such, one might consider a "normalized" version of $\pi(s|\cdot)$, which we call $\tilde{\pi}(s|\omega)$, which generates posterior $\mu = \mu(\cdot|s)$. $\tilde{\pi}(s)$ is the highest possible value of $\pi(s|\cdot)$ that yields this posterior given μ_0. One can then express $\pi(s|\omega) = \alpha(s)\tilde{\pi}(s|\omega)$, where $\alpha(s) \in [0, 1]$ scales the probability that $\mu(\cdot|s)$ is realized.[3] One would then write the cost of an experiment as α-adjusted cost of each such signal, i.e.

$$c(\pi) = \int_{\mathcal{S}} \tilde{c}(\tilde{\pi}(s))d\alpha(s)$$

The appeal of such a representation is similar to that of the posterior-separable measure of uncertainty, in that a DM can optimize signal-by-signal. However, this representation has the additional advantage that it is defined without reference to a particular prior. Since the ordering over Π is also defined without reference to a particular prior, this seems like a more natural way to model information. We will also show that it has advantages when modeling information acquisition in games.

Even though the cost has a separable representation, one still must ensure that the total cost is increasing in informativeness. Thus, for any three signals s_1, s_2, s_3,

$$\tilde{c}(\tilde{\pi}(s_1))\alpha(s_1) + \tilde{c}(\tilde{\pi}(s_2))\alpha(s_2) > \tilde{c}(\tilde{\pi}(s_3))\alpha(s_3)$$

whenever $\alpha(s_1)\tilde{\pi}(s_1) + \alpha(s_2)\tilde{\pi}(s_2) = \alpha(s_3)\tilde{\pi}(s_3)$, i.e. one replaces signals s_1 and s_2 with a merged signal s_3.

As in the previous subsections, we ask whether the representation of $\tilde{c}$ is essentially unique. I say that it is so if one can transform any $\tilde{c}'$ that represents the same cost can be transformed into $\tilde{c}$ by looking at N constants $\{\lambda_i\}_{i=0}^{N-1}$. The precise relationship of these constants to $\tilde{c}$ is somewhat complicated, and so is described in detail in Appendix B when this representation is constructed.

[3]Note that α is not a probability measure, since it is possible that $\int_{\mathcal{S}} d\alpha(s) \geq 1$. This will be made clearer in the following section.

3 Main Results

3.1 Representation Theorem

In the previous section, I provided five different ways of representing information. Surprisingly, these are all equivalent.

Theorem 1: *The following are equivalent:*

1. *There is an binary relation* $\succsim$ *over* Π *that satisfies Axioms (1)-(6);*
2. *There is an essentially unique cost function c that is strictly increasing in the Blackwell order and satisfies (IR);*
3. *Given prior* μ_0*, there is an essentially unique utility function u (defined almost everywhere) which the DM maximizes contingent on his posteriors, and is strictly increasing in the Blackwell order;*
4. *Given prior* μ_0*, there is an essentially unique posterior-separable cost of information by measure of uncertainty H;*
5. *There is an essentially unique separable cost of signals* $\tilde{c}$ *that is strictly increasing in the Blackwell order.*

While I relegate the formal details to Appendix B, I provide a sketch of the main arguments here. Some of the directions are unsurprising. For instance, that representations (2)-(5) each imply (1) is immediate. That (1) implies (2) is also unsurprising, since $\succsim$ satisfies the vNM axioms; though the set of possible signals is infinite, this is taken care of by Belief Continuity (Axiom (4)).

More surprising is that (2) implies (4). As it is not so clear why this is true, I spend more time on this than the other directions. The proof is constructive; I here illustrate an example where $N = 3$.

Starting with prior μ_0, assign $H(\mu_0) = 0$, and assign arbitrary values of $H(\delta_1), H(\delta_2)$, where δ_n is the Dirac measure on state ω_n. Letting the fully informative experiment be π_∞, (illustrated in Figure 1), this pins down the value of $H(\delta_3)$, since

$$c(\pi_\infty) = \sum_{n=1}^{3} \mu_0(\omega_n) H(\delta_n)$$

Next, consider the set of possible experiments that generate some posterior μ with positive probability. Given μ_0, μ is determined by the ratio of probabilities π that s is generated

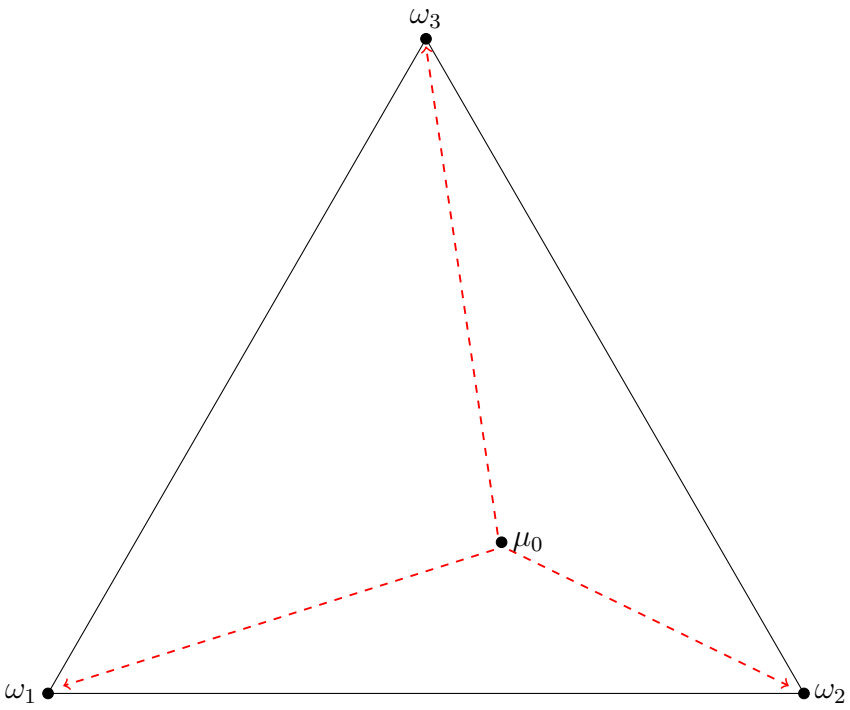

Figure 1: Fully informative experiment π_∞

given $\omega \in \Omega$. To generate μ with the highest probability possible, one increases π as much as possible, while maintaining the same ratio. This occurs where $\pi(s|\omega^*) = 1$ for some ω^*. One such possible experiment is where, if μ is not generated, then the state is fully revealed. Such an experiment is illustrated in Figure 2, where $\omega^* = \omega_3$. Since we already know the values of H for the other posteriors, as well as the cost of the entire experiment, this provides a value for $H(\mu)$ as well.

One still must ensure that the values of $H(\mu)$ found in this manner are consistent with each other for all possible experiments that generate μ. Consider some experiment π which generates posteriors as in Figure 3. I show that this cost is pinned down by that of the experiment in Figure 2.

Consider the following alternative experiment: for each μ generated in the experiment in Figure 3, randomize between running the analogous experiment in Figure 2, and revealing all information. By (IR) this can be thought of as deterministically running a particular experiment (illustrated in Figure 4) that generates the same posteriors with the same probabilities τ. In Appendix B, I show that for the right choice of randomizing probabilities, this can also be viewed as randomizing between π and π_∞. We already know the values of c for all these experiments, and that for all these experiments other than π, the cost is pinned down by the expected posterior value of H. For equality of costs to hold, as it must by (IR), the cost of

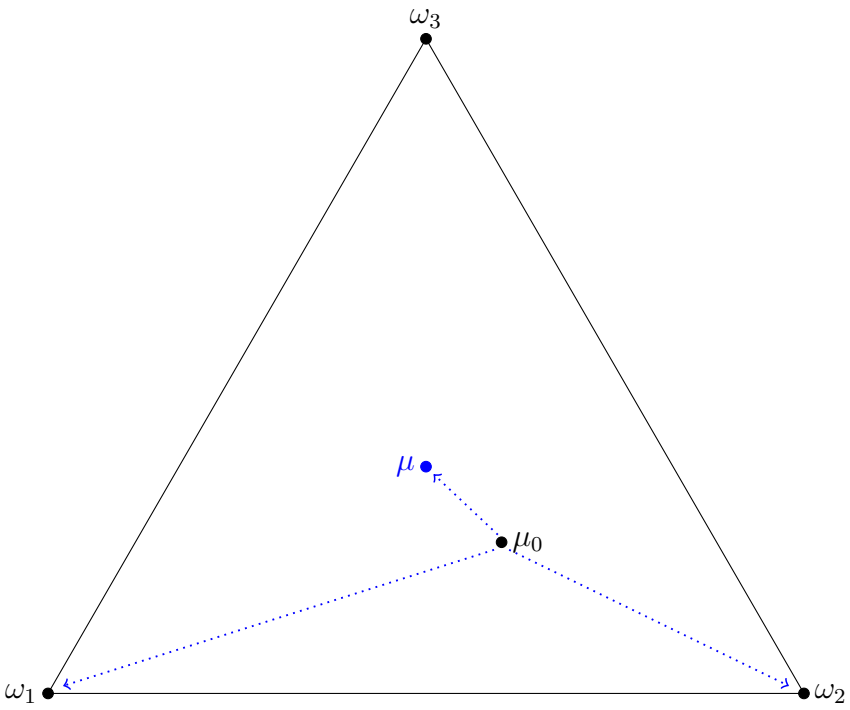

Figure 2: Experiment π^*_μ

π must also be given by the expected posterior value of H.

To round out the sketch of the proof, one can see that (4) implies (2) by standard convex analysis/duality theory results. For any μ, we can think of $H(\mu)$ as being the expected utility of a DM who optimizes at μ, as long as we can find an appropriate utility function u for the DM. This is done by use of supporting hyperplanes at each μ, which can be thought of as defining $u(x^*(\mu), \omega)$, where $x^*(\mu) = \arg\max_{x \in X} \sum_{\omega \in \Omega} u(x, \omega)\mu(\omega)$. I illustrate this in a two-state example in Figure 5.

Lastly, to see that (4) implies (5), recall that the determination of the posterior μ given μ_0 depends only on the ratios between $\pi(s|\omega)$ and $\pi(s|\omega')$ for any $\omega, \omega' \in \Omega$. As we saw in the construction of $H(\cdot)$, there is some state ω^* for which this ratio is maximal. One can then define the measure $\alpha(s)$ relative to the maximum probability that this signal is generated, for which $\pi(s|\omega^*)$ would be 1. Thus, for any signal s that is generated with positive probability, $\alpha(s) = \pi(s|\omega^*)$, and so $\tilde{c}(\tilde{\pi}(s)) = \frac{H(\mu)}{\mu(\omega^*)/\mu_0(\omega^*)}$. In Appendix B, I formally derive the details of this representation, and show that it does not depend on μ_0.

3.2 Comparative Statics

The representation of information in Theorem 1 allows for comparative statics across decision makers regarding preferences over information. As seen in Theorem 1, any cost

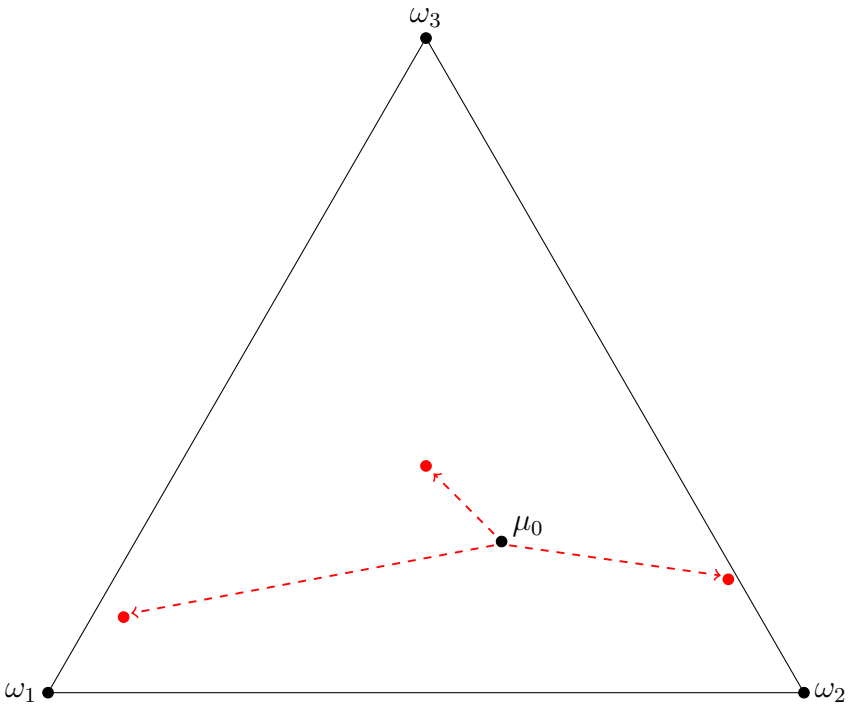

Figure 3: Experiment π

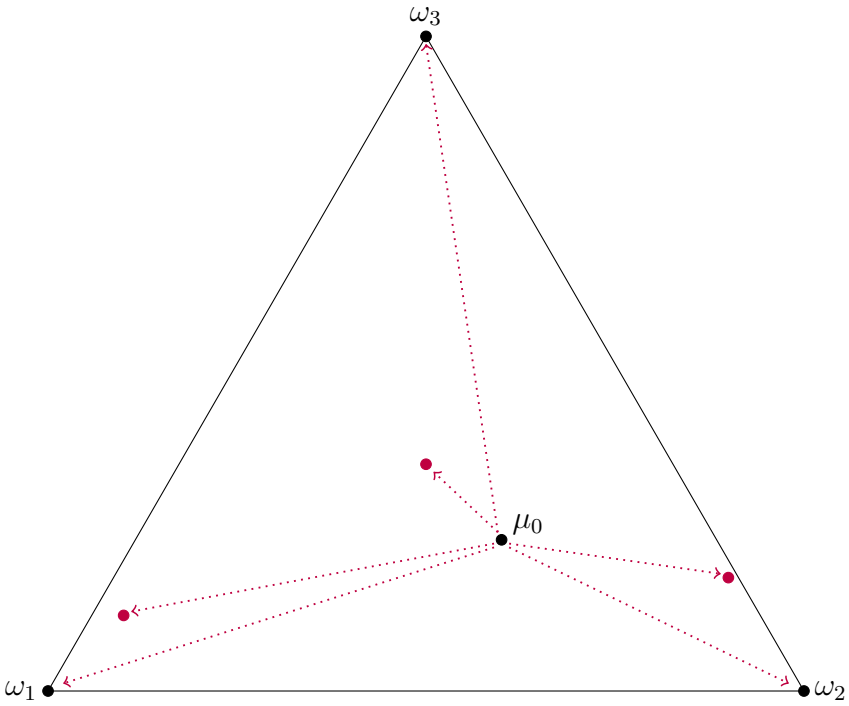

Figure 4: Experiment $\hat{\pi}$, which pins down $c(\pi)$

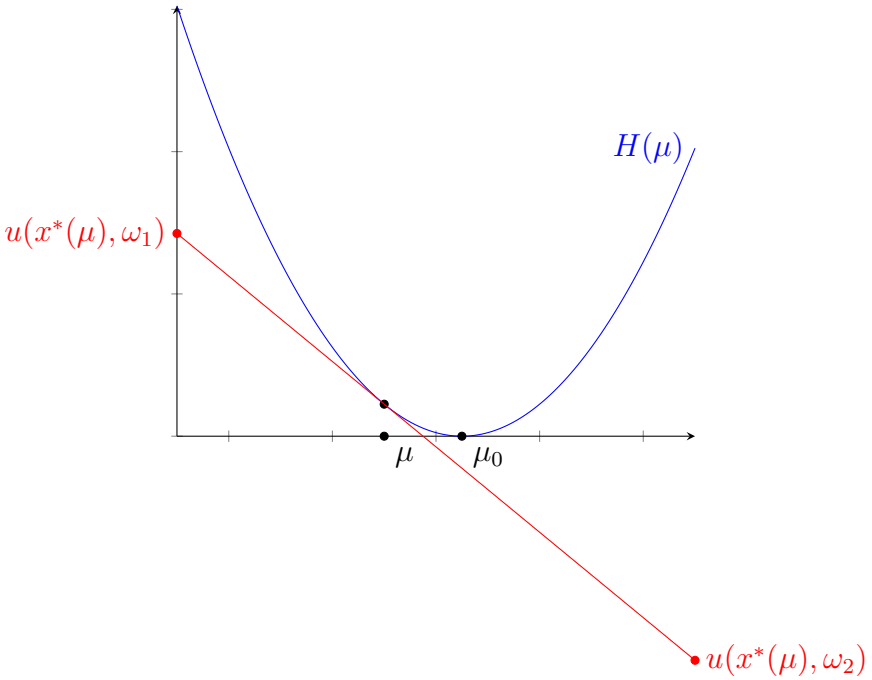

Figure 5: Construction of u

function c for acquiring information can, if it satisfies the (IR) condition, be represented by some posterior-separable measure of uncertainty H. This allows a comparison of how H depends on c. Specifically, one can ask how H changes as information acquisition becomes more costly. Thus, DM 1 is said to have *higher marginal costs of information acquisition* than DM 2 if, for any π' that is Blackwell more informative than π,

$$c_1(\pi') - c_1(\pi) \geq c_2(\pi') - c_2(\pi)$$

Corollary 2: *DM 1 has higher marginal costs of information acquisition than DM 2 if and only if $H_1 \equiv H_2 + \tilde{H}$ for some convex function $\tilde{H}$.*

4 Implications for Rational Inattention

There has been a large literature in recent years that assumes that the cost of information acquisition is proportional to *mutual information*, i.e. entropy reduction. In many cases, the literature has considered this cost for problems with arbitrary distributions of states, or at the very least, a family of distributions. Thus, Mackowiak and Wiederholt (2009) represent information costs by entropy reduction for a sticky-prices model in which the distribution of shocks is normally distributed (but any variance is possible). Cabrales et al. (2013) show

that mutual information is the only cost function that, regardless of prior, ranks information for an investment problem. Matejka and McKay (2015) use mutual information as a basis for the multinomial logit, showing that whenever the values for different choices are independent of each other, the distribution of choices conditional on the state is the multinomial logit. Ravid (2017) examines a rationally inattentive bargainer who acquires information about the distribution of offers from potential sellers; since the distribution of offers is endogenous, the model assumes mutual information costs for any potential distribution of offers. Denti (2017) constructs a model of information acquisition about one's opponent's information, which is also endogenous. For many of his applications, he assumes that information acquisition costs are proportional to mutual information. Mensch (2017) shows that a DM acquires signals ordered by the monotone likelihood ratio property for any decision problem with increasing differences and any prior if and only if his costs are proportional to mutual information. Given the vast use of rational inattention in recent years, it is important to ask what the results here imply for this approach.

At first glance, the results of the previous section may appear to be good news for the rational inattention literature. Theorem 1 shows that, as is often assumed, there exists a posterior-separable measure of uncertainty H that represents a cost of information acquisition whenever Axioms 1-6 are satisfied. As informational entropy is an example of such a function H, one may think that Theorem 1 justifies the use of entropy reduction as a proxy for information costs. I will argue that such an approach, while possibly justifiable for a single decision maker with a fixed prior, is problematic in dynamic environments and in games. I will illustrate this in the following subsection by examining an application to a principal-agent problem with moral hazard.

4.1 An example: moral hazard with endogenous monitoring

Consider a principal (P) who wants to induce an agent (A) to exert costly effort. For simplicity, I assume effort is binary, $e \in \{0, 1\}$. The principal cannot directly observe whether the agent has exerted effort, but can acquire information about this through a (costly) monitoring system π_P. The principal may make transfer $t(s)$ contingent on each signal s induced by π. Assume limited liability, i.e. $t \geq 0$. Effort is costly for the agent, with cost of effort

$$c(e) = \begin{cases} 0, & e = 0 \\ c_A < 1, & e = 1 \end{cases}$$

The order of moves are as follows: (1) the principal chooses a monitoring structure π and a contingent transfer scheme t;[4] (2) The agent chooses to exert effort (possibly randomizing) via strategy σ; (3) a signal s is generated contingent on the (distribution of) effort chosen by the agent, and the monitoring structure π; and (4) the transfer t is made. I assume that the total payoff for each player is

$$v_P(\pi, t, e) = e - c(\pi) - \sum_{e=1,2} \sigma_A(e) \int t(s) d\pi(s|e)$$

$$v_A(\pi, t, e) = -c_A + \sum_{e=1,2} \sigma_A(e) \int u_A(t(s)) d\pi(s|e)$$

where u_A is some concave function (i.e. the agent is risk-averse over monetary outcomes). We assume that the principal-optimal equilibrium is implemented.

Using the approach of the rational inattention literature, the cost of information acquisition via signal π is

$$c(\pi) = E_{<\mu|\pi>}[H(\mu)] - H(\mu_0)]$$

where $\mu_0 = \sigma$, i.e. the distribution of effort levels exerted by the agent. Perhaps somewhat surprisingly, this modeling choice enables the principal to achieve the first-best solution. Suppose that the principal implements the binary signal structure $\mathcal{S} = \{s_0, s_1\}$, where

$$\pi(s_0|\ e) = \begin{cases} 1, & e = 0 \\ 0, & e = 1 \end{cases}$$

and gives transfers

$$t(s) = \begin{cases} c_A, & s = 1 \\ 0, & s = 0 \end{cases}$$

Under this signal structure, the agent has a weak incentive to exert effort, as effort is perfectly observable. Now let us check the cost of information acquisition conditional on this choice of effort. The prior μ_0 places probability 1 on the belief that the agent has exerted effort; therefore, with probability 1, signal s_1 is generated, which induces the same posterior. Since $\mu(\cdot|s_1) = \mu_0$, we conclude that $c(\pi) = 0$. Of course, this undermines the entire point of modeling endogenous monitoring, which is to say that better monitoring *should* be (more) costly. This *reductio ad absurdum* argument should indicate that the rational inattention approach is incorrect in a game-theoretic environment, where the distribution of states depends on the

[4]Note that I assume that this is chosen before the agent chooses whether to exert effort, as if not, the effort is sunk, and so there is no point to monitoring.

action of some other player.

The approach espoused by Theorem 1 avoids this issue. Note that the theorem states that one can represent the cost of information acquisition by some measure of uncertainty H for *any* prior distribution μ_0. Thus one can assume any reference distribution μ^* as a prior, and consider the resultant function H with reference to that distribution, even if it is not the true distribution of effort. Of course, this is only for the purpose of analyzing the costs; for the benefits, one needs to use the true distribution. The tradeoff between better incentives from monitoring and costs of information acquisition can be then made with respect to that reference distribution. I illustrate this more explicitly in the analysis below.

The following observation simplifies the ensuing analysis.

Lemma 3: *There exists an optimal signal structure* $(\pi, \mathcal{S})$ *such that* $|\mathcal{S}| = 2$.

One can now express $\mathcal{S} = \{s_0, s_1\}$, and optimize the monitoring structure with respect to this. If it is too costly to monitor, one simply sets $\frac{\pi(s_0|e=1)}{\pi(s_0|e=0)} = \frac{\pi(s_1|e=1)}{\pi(s_1|e=0)}$ and offers $t = 0$. If monitoring is optimal, the principal minimizes the total costs of monitoring and transfers. Since we are looking at the principal-optimal solution, the agent exerts effort with probability 1. WLOG, assume that $\frac{\pi(s_0|e=1)}{\pi(s_0|e=0)} < \frac{\pi(s_1|e=1)}{\pi(s_1|e=0)}$. The incentive compatibility constraint (IC)

$$-c_A+\pi(s_1|e=1)u_A(t(s_1))+\pi(s_0|e=1)u_A(t(s_0)) \geq \pi(s_1|e=0)u_A(t(s_1))+\pi(s_0|e=0)u_A(t(s_0)) \tag{3}$$

must be binding, which implies that $t(s_1) > t(s_0)$.

Assume now that H is differentiable.[5] Then the principal's problem is

$$\min_{t(s_1),t(s_0)} \pi(s_1|e=1)t(s_1) + \pi(s_0|e=1)t(s_0)$$

$$\text{s.t.}\,(IC), (IR)$$

Turning to the monitoring problem, we look at the marginal costs from changing the signal structure. Recall that by Bayes' rule, if we use μ^* as our reference prior,

$$\mu(e|s) = \frac{\mu^*(\omega)\pi(s|e)}{\tau_\pi(s)}$$

$$\tau_\pi(s) = \mu^*(e=1)\pi(s|e=1) + \mu^*(e=0)\pi(s|e=0)$$

Therefore, one can write the marginal cost of changing the signal structure, by (say) increasing $\pi(s_1|e=1)$ and decreasing $\pi(s_0|e=1)$, as (again, with all beliefs generated with respect to

[5]By Rockafellar, Theorem 25.5, one can approximate any convex or concave function by one that is also differentiable.

μ^* given π)

$$\frac{\partial c}{\partial \pi(s_1|1)}(\pi) = \mu^*(1)[H(\mu(\cdot|s_1)) - H(\mu(\cdot|s_0))] \tag{4}$$

$$+\frac{\mu^*(1)(1-\mu^*(1))}{\tau_\pi(s_1)}[\frac{\partial H}{\partial \mu(1)}(\mu(\cdot|s_1))\pi(s_1|0) - \frac{\partial H}{\partial \mu(0)}(\mu(\cdot|s_1))\pi(s_1|1)]$$

$$-\frac{\mu^*(1)(1-\mu^*(1))}{\tau_\pi(s_0)}[\frac{\partial H}{\partial \mu(1)}(\mu(\cdot|s_0))\pi(s_0|0) - \frac{\partial H}{\partial \mu(0)}(\mu(\cdot|s_0))\pi(s_0|1)]$$

At the optimum, this must be equal to the marginal benefit, which is the ability to motivate the agent to exert effort while paying him less. The precise marginal benefit is, by the envelope theorem, exactly $t(s_1)$. Thus, at the principal-optimal solution, we set $t(s_1)$ equal to (4). Analogously, for the optimization with respect to $e = 0$, as $e = 0$ does not appear in the objective for the principal, the analogous expression of (4) (with $e = 1$ replaced with $e = 0$) must equal 0.

I summarize the first-order conditions of the solution of the principal's problem in the following proposition.

Proposition 3: *The solution to the principal's monitoring problem, if effort is optimal, satisfies*

$$\frac{\partial c}{\partial \pi(s_1|1)}(\pi) = t(s_1)$$

$$\frac{\partial c}{\partial \pi(s_1|0)}(\pi) = 0$$

While the general expression of (4) is complicated, two points are worth making here. First, for specific parameterizations of H and u_A, the expressions will simplify, and allow for an analytic solution. Second, it is clear that the fully-informative experiment is not necessarily optimal anymore, since if the marginal cost is too high relative to the $t(s_1)$ (which must be precisely c_A at the fully informative signal), then it would be beneficial to reduce $\pi(s_1|e = 1)$. Thus the model of information in this paper successfully captures the potential for flexible monitoring, something which rational inattention, as currently used, cannot do.

The principal-agent model also highlights the usefulness of the "cost of signals" representation. Notice that the optimal signal structure is binary; s_1 is more likely to be realized for $e = 1$, while s_0 is more likely for $e = 0$. In other words, $\omega^*(s_1) = 1$, and $\omega^*(s_0) = 0$; this, in turn, implies that $\alpha(s_i) = \pi(s_i|e = i)$. A sufficient statistic for $\tilde{\pi}(\tilde{s}_1)$ is

$$\frac{\pi(s_1|e=1)}{\pi(s_1|e=0)} = \frac{\alpha(s_1)}{1-\alpha(s_0)}$$

Thus let (with slight abuse of notation)

$$\tilde{\pi}(s) = \begin{cases} \frac{\alpha(s_1)}{1-\alpha(s_0)}, & s = s_1 \\ \frac{1-\alpha(s_1)}{\alpha(s_0)}, & s = s_0 \end{cases}$$

Note that $\tilde{\pi}(s_1) > 1 > \tilde{\pi}(s_0)$. The first-order condition (4) can now be expressed as

$$t(s_1) = \tilde{c}(\tilde{\pi}(s_1)) + \frac{\partial \tilde{c}}{\partial \tilde{\pi}}(\tilde{\pi}(s_1))\frac{\alpha(s_1)}{1-\alpha(s_0)} - \frac{\partial \tilde{c}}{\partial \tilde{\pi}}(\tilde{\pi}(s_0))\frac{\alpha(s_1)}{\alpha(s_0)} \tag{5}$$

which is much simpler.

4.2 Prior independence

In the previous subsection, I showed that the standard rational inattention approach leads to the unacceptable conclusion that monitoring is costless in a principal-agent problem. What goes wrong?

The problem with using mutual information as a measure of uncertainty is that it is *prior-dependent*: that is, the cost of information acquisition as measured by the expected reduction of entropy depends, for a given experiment, on the prior that one uses.[6] This led to the ability of the principal to exploit this property to fully reveal the effort level of the agent. By contrast, intuitively, costly information acquisition should be *prior independent*; the cost of acquiring signal π should not depend on the prior one has.

To see why this should be the case, consider the following scenario. Suppose that an environmental researcher is attempting to measure the level of carbon dioxide in the atmosphere. At her disposal will be various instruments that can measure carbon dioxide; however, there will be some noise in the measurement, and so the experiments will not necessarily be fully informative. The researcher will have some prior about the level of carbon dioxide, which will inform her decision regarding which (and how many) experiments to run. Intuitively, if the prior of the researcher were different, the set of experiments does not change; hence the cost of running a particular experiment should remain the same (though the *choice of which* experiment to run may change as a result).

This intuition is even stronger in an extensive-form game, in which the prior that one player has depends on the actions of his opponents (as was the case in the principal-agent model above). Suppose that one's opponent were to deviate, and so one's prior would then

[6]This has already been noted in the case of mutual information by Cabrales et al. (2013). The analysis here goes beyond theirs to consider all possible H functions.

change as well. The player has no way to detect this deviation before acquiring information. Hence the decision calculus for the player to acquire information should remain the same.

Nevertheless, a frequent justification for the use of mutual information is that the costs are cognitive: there is a limited amount of data that the DM can process at any given moment. Shannon (1948) famously showed that informational entropy is the correct measure of efficient flow of data through a limited-capacity channel. Thus, one may think that as a result, mutual information is the correct measure of rational inattention costs.

Yet even here, one could argue that the problem faced here is one of experimentation, in the form of "thought experiments." It may be hard for the DM to reason through exactly what is going on, and consider all of the facts. So, the DM may consider only certain hypotheticals, some of which may be harder to think about than others. He may then think as follows: "If the state ω were the true state, then I should expect to get answer s to the hypothetical with some probability $\pi(s|\omega)$, and similarly for other possible states. Since I think that the answer the the hypothetical is s, that is how I will assess the state."

Additionally, as I will argue in the next subsection, I believe that there is some confusion in the literature regarding the interpretation of Shannon's results. I will discuss there the distinctions in the usage of the terms "experiment" and "information" as modeled by Blackwell and Shannon, respectively, which I believe is the source of the confusion.

A natural question at this point is whether there is any function H that satisfies the desired prior-independence property. It is clear from the previous subsection that this cannot work for any non-zero cost that satisfies axioms (1)-(6), since any such H will result in a cost of 0 when the prior is a Dirac delta function on a particular state, no matter what the experiment. By continuity arguments, the cost will be approximately 0 for nearby priors in the interior of the simplex as well. One may try to relax the constraint by considering H that satisfies the axioms in the interior of the simplex, so as to avoid this particular issue. This could allow for infinite values of H at the boundary, which would make the cost undefined as measured there, but consistent with a prior-independent cost function on the interior by taking limits. Yet even with this relaxation, the result is negative: no such H exists.

Proposition 4: *There exist no non-zero functions* $c : \Pi \to \mathbb{R}$ *and* $H : int(\Delta(\Omega)) \to \mathbb{R}$ *such that for all* $\mu_0 \in int(\Delta(\Omega))$*,*

$$c(\pi) = E_{<\mu|\pi>}[H(\mu)] - H(\mu_0)$$

Thus the critique from the principal-agent problem is robust to the specification of cost function.

4.3 Shannon information vs. Blackwell information

At this point, one is still left with the puzzle of why the use of mutual information leads to such perverse results. After all, Shannon's definition of information is mathematically correct in describing the amount of information that can flow through a limited channel, and has proven very useful across a wide variety of disciplines, including information theory, computer science, and statistics. Yet clearly, it leads to unacceptable conclusions in the principal-agent game. In the previous section, I discussed how mutual information failed the prior independence property; how is this consistent with the widespread usefulness of defining information by entropy?

It appears to me that there are two different definitions of information that are in use, and that they have been conflated in the rational inattention literature. Economic theorists have tended to use the definition of information found in Blackwell (1951, 1953), in which there is some underlying unknown state ω. Distinct from this, there is some signal s which one observes that is correlated with ω. One then updates one's beliefs by Bayes' rule.

Examining information as defined by Shannon (1948), it becomes clear that the information conveyed by a signal is about *the signal itself*. Information is measured by a bit, that is, a signal whose realization is 0 or 1. This signal is not the result of an experiment from some underlying state of nature ω; rather, it noiselessly represents, say, a letter/character that is being transmitted through some channel. The character may be generated randomly through some Markov process, but it does not represent some additional, hidden underlying variable. In this way, Shannon information does not distinguish between ω and s.

This definition is the one that he uses when axiomatizing a measure of information (Section 6 of Shannon, 1948). At first, it may seem that he is examining the same definition of information as Blackwell, as he is considering a set of possible events $i \in \{1, ..., n\}$ that occur with respective probabilities p_i. However, his Axiom 3, that of additivity, makes it clear that these probabilities are also being used to define experiments π. Formally, his Axiom 3 states the following. Suppose that one can express some experiment π as comprising two sub-experiments, π_1 and π_2, such that, conditional on the realization of certain signals $s \in \mathcal{S}' \subset \mathcal{S}$, experiment π_2 is then run. Then

$$H(\pi) = H(\pi_1) + \sum_{s \in \mathcal{S}'} \pi_1(s) H(\pi_2) \tag{6}$$

This definition is appropriate for the environment that Shannon is describing, since he is considering transmission of characters. Thus, the information being provided is identical to the signal. So, if one sends over some bits representing $s \in \mathcal{S}$, and conditional on $s \in$

$\mathcal{S}'$, one then sends additional bits, it would make sense to represent information as the expected total number of bits that will be sent in this message. This is the idea behind why mutual information $I(X,Y)$, as defined by Kullback and Leibler (1951), is commutative, i.e. $I(X,Y) = I(Y,X)$: it expresses amount of information that is shared between X and Y, as measured in the number of bits shared by the expression of X and Y. Thus, writing an experiment in terms of the mutual information between $\mathcal{S}$ and Ω is equal to that shared between Ω and $\mathcal{S}$.

Using the definition in (6), if one were to describe the total entropy from the joint distribution of (s,ω) given π, it would be

$$\begin{aligned}
&\sum_{\omega\in\Omega}\mu_0(\omega)\ln\mu_0(\omega) + \sum_{\omega\in\Omega}\sum_{s\in\mathcal{S}}\mu_0(\omega)\pi(s|\omega)\ln(\pi(s|\omega)) \\
&= \sum_{s\in\mathcal{S}}\tau_\pi(s)\ln(\tau_\pi(s)) + \sum_{s\in\mathcal{S}}\tau_\pi(s)\mu(\omega|s)\ln(\mu(\omega|s)) \qquad (7)
\end{aligned}$$

So, the mutual information (i.e. the shared bits) will be

$$\begin{aligned}
I(\mathcal{S},\Omega) &= \sum_{\omega\in\Omega}\mu_0(\omega)\ln\mu_0(\omega) - \sum_{s\in\mathcal{S}}\tau_\pi(s)\mu(\omega|s)\ln(\mu(\omega|s)) \\
&= \sum_{s\in\mathcal{S}}\tau_\pi(s)\ln(\tau_\pi(s)) - \sum_{\omega\in\Omega}\sum_{s\in\mathcal{S}}\mu_0(\omega)\pi(s|\omega)\ln(\pi(s|\omega)) \qquad (8)
\end{aligned}$$

Equality holds because mutual information is commutative: for any two random variables X, Y,

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Therefore, as seen, s and ω play symmetric roles.

This definition, however, is distinct from Blackwell's definition of information, in which the experiment is defined *asymmetrically*. That is, in Blackwell's definition, the outcome s from experiment π is contingent on some underlying state ω; this definition does not depend on a prior. However, it would be meaningless to describe ω in terms of s without reference to a prior.

As a final remark, it is possible that this distinction has not always been noticed because, whenever a signal is Blackwell more informative, the informational entropy as expressed by (2) decreases. This is because the result of a Blackwell experiment s is correlated with state ω. Thus there will be some mutual information between the two variables. Moreover, since entropy is positive and concave, any Blackwell more informative experiment will have higher entropy reduction. Of course, as shown above, as well as by Cabrales et al., the exact

difference in entropy will depend on the prior, since the amount of information "shared" by the experiment with the prior will depend on the prior. So, while entropy reduction satisfies the desired property of monotonicity in Blackwell informativeness, it does so for a slightly different reason, since Blackwell and Shannon information have different bases.[7]

5 Conclusion

This paper provides a general treatment of information, connecting it to the various other ways that information has been represented in the past. Using an axiomatic framework, I showed the equivalence between many of these representations. Making these connections explicit serves to clarify the relationships between different information structures. In particular, I showed that the "indifference to randomization" property (IR) implies independence between the signal realizations: that is, one can treat each signal realization separately from the rest of the signal structure. Thus, if one aims to optimize one's information acquisition, one can look at tradeoffs between pairs of signal realizations while ignoring the others.

I also showed that the representation here has implications for the literature on rational inattention. We saw that using entropy reduction as a proxy for information acquisition can lead to perverse conclusions. In exploring this, I showed that entropy reduction violates the desired property of prior independence, and that informative experiments as defined by the axiomatization of entropy are not identical to Blackwell experiments.

While the results do not imply that one should not use entropy reduction as a proxy for information, they do provide a strong caveat. One must be wary when using entropy reduction across multiple potential priors, as they do not accurately measure the effects and incentives of information acquisition. Using the representations here (either using a measure of uncertainty with respect to a reference prior, or by looking at costs of individual signal realizations), one can keep the cost consistent across multiple potential priors.

References

Blackwell, D. (1951): "Comparison of experiments." *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, 93-102.

[7]It is also possible that some confusion may have arisen about additivity, in that if one runs two experiments π_1, π_2 in succession, then the total reduction in entropy is equal to the reduction due to π_1 from going from μ_0 to μ_1, plus the reduction due π_2 from going from μ_1 to μ_2. Yet such "additivity" is present for any convex function H, and is distinct from the additivity in entropy described by Shannon's Axiom 3.

——— (1953): "Equivalent comparisons of experiments." *Annals of Mathematical Statistics*, 24, 265-272.

Cabrales, A., Gossner, O., and Serrano, R. (2013): "Entropy and the value of information for investors." *American Economic Review*, 103, 360-377.

Caplin, A., and Dean, M. (2013): "Behavioral implications of rational inattention with Shannon entropy." NBER working paper No. 19318.

——— (2014): "Revealed preference, rational inattention, and costly information acquisition." Working paper.

——— (2015): "Revealed preference, rational inattention, and costly information acquisition." *American Economic Review*, 105, 2183-2203.

Caplin, A., Dean, M., and Leahy, J. (2017): "Rationally inattentive behavior: Characterizing and generalizing Shannon entropy." NBER Working Paper No. 23652.

Cover, T. M., and Thomas, J. A. (2006): *Elements of Information Theory.* Hoboken, NJ: John Wiley & Sons, Inc.

De Oliveira, H. (2014): "Axiomatic foundations for entropic costs of attention." Working paper.

De Oliveira, H., Denti, T., Mihm, M., and Ozbek, K. (2016): "Rationally inattentive preferences and hidden information costs." *Theoretical Economics*, 12, 621-654.

Debreu, G. (1954): "Representation of a preference ordering by a numerical function." Cowles Foundation Discussion Paper No. 97.

Dekel, E., Lipman, B., and Rustichini, A. (2001): "Representing preferences with a unique subjective state space." *Econometrica*, 69, 891-934.

Denti, T. (2016): "Unrestricted information acquisition." Working paper.

Ely, J., Fraenkel, A., and Kamenica, E. (2015): "Suspense and surprise." *Journal of Political Economy*, 123, 215-260.

Ergin, H. and Sarver, T. (2010): "A unique costly contemplation representation." *Econometrica*, 78, 1285-1339.

Kamenica, E., and Gentzkow, M. (2011): "Bayesian persuasion." *American Economic Review*, 101, 2590-2615.

——— (2014): "Costly persuasion." *American Economic Review (Papers and Proceedings)*, 104, 457-462.

Kreps, D. (1979): "A representation theorem for 'preference for flexibility.'" *Econometrica*, 47, 565-577.

Kullback, S. and Leibler, R. A. (1951): "On information and sufficiency." *Annals of*

Mathematical Statistics, 22, 79-86.

Mackowiak, B. and Wiederholt, M. (2009): "Optimal sticky prices under rational inattention." *American Economic Review*, 99, 769-803.

Matejka, F., and McKay, F. (2015): "Rational inattention to discrete choices: A new foundation for the multinomial logit model." *American Economic Review*, 105, 272-298.

Mensch, J. (2017): "Rational inattention and the monotone likelihood ratio property." Working paper.

von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton, NJ. Princeton University Press, 1947.

Ravid, D. (2017): "Bargaining with rational inattention." Working paper.

Rockafellar, R. T. *Convex Analysis*. Princeton, NJ. Princeton University Press, 1970.

Shannon, C. (1948): "A mathematical theory of communication." *Bell System Technical Journal*, 27.

Sims, C. (1998): "Stickiness." *Carnegie-Rochester conference series on public policy*, 49, 317-356.

——— (2003): "Implications of rational inattention." *Journal of Monetary Economics*, 50, 665-690.

Woodford, M. (2008): "Inattention as a source of randomized discrete adjustment." Working paper.

Yang, M. (2015): "Coordination with flexible information acquisition." *Journal of Economic Theory*, 158, 721-738.

A Topology of Experiments

A naive way to model the topology of experiments Π is to endow it with the product weak* topology state-by-state. However, the set of experiments which lead to meaningful posteriors is then not compact. This is because there is the possibility that the limit of a sequence of measures has a singular component for which Bayes' rule would not be well-defined.

In order to circumvent this issue and properly define a topology over experiments π, we do so with respect to some reference prior μ_0. Endow $\Delta(\Omega)$ with the Euclidean topology, and $\Delta(\Delta(\Omega))$ with the weak* topology. The latter is known to be metrizable, and so is a Polish space. Let δ be a metric on $\Delta(\Delta(\Omega))$. I define the distance between π and π' to be $\delta(\tau_\pi, \tau_{\pi'})$ given μ_0. Using this metric, Π is compact. Assign Π the Borel σ-algebra for the purpose of randomizing over experiments.

I now present the proofs of Lemmas 1 and 2, using these topologies.

Proof of Lemma 1: Given τ_π and μ_0, for any measurable set $A \subset \Delta(\Omega)$, let

$$\pi(A|\omega) \equiv \frac{\tau_\pi(A) \int_A \mu(\omega) d\tau_\pi(\mu)}{\mu_0(\omega)} \tag{9}$$

as must be the case by Bayes' rule. By the Lebesgue differentiation theorem, there exists a π which defines the measure over $\mathcal{S}$ that satisfies the above equation, which is unique up to a set of Lebesgue measure 0. Since $|\Omega|$ is finite, one can uniquely aggregate $\pi(\cdot|\cdot)$ state-by-state, up to measure 0. $\square$

Note that this approach in showing an equivalence between π and τ_π in Lemma 1 avoids the pitfalls of the naive approach, since we eliminate the possibility of having a singular measure, unless all states have a singular measure on the same signal s. Moreover, they will generate the singular s at the same relative likelihood, as defined by the limits of open sets around s (as seen by the derivation via the Lebesgue differentiation theorem). Hence one can extend Bayes' rule even to singular signal realizations.

The isomorphism between π and τ_π with respect to μ_0 provided in Lemma 1 will be exploited in the proof of Lemma 2, presented below.

Proof of Lemma 2: Since $\Delta(\Delta(\Omega))$ is a Polish space, compactness on $\Delta(\Delta(\Omega))$ is equivalent to sequential compactness. Thus consider a sequence of experiments $\{\pi_k\}_{k=1}^\infty$ such that $\lim_{k\to\infty} \pi_k$ exists and is more informative than π given μ_0. One can label this limit by π_∞. Since the Blackwell order is independent of the prior, we are done as long as the limit π_∞ is the same regardless of the prior.

To demonstrate this, for any measurable set $A \subset \mathcal{S}$, define $\pi_k(A|\omega)$ as in (9).[8] Then $\pi_k(\cdot|\omega)$ converges in the weak* topology because τ_{π_k} does. Thus, with respect to μ_0', the limit of τ_{π_k} is well-defined as

$$\lim_{k\to\infty} \tau_{\pi_k} = \lim_{k\to\infty} \sum_{\omega\in\Omega} \pi_k(\cdot|\omega)\mu_0'(\omega)$$

where all limits are in the weak* topology. The upper- and lower-contour sets then describe the same sets of experiments regardless of the prior. □

B Proofs of Theorems

B.1 Proof of Theorem 1

I show that (1) $\implies$ (2), (2) $\implies$ (4), (4) $\implies$ (5), and (4) $\implies$ (3). That (3) $\implies$ (1) is trivial, since in the former one just examines which experiments generate higher expected utility, which must continuous in beliefs by the Berge maximum theorem, and satisfies (IR) because the expected utility depends only on the distribution of posteriors and the decision made at those posteriors. Similarly that (5) $\implies$ (1) is trivial since $\tilde{c}$ is continuous in the normalized probabilities $\tilde{\pi}$, and (IR) is satisfied because the cost depends only on the probabilities that signal s with the given $\tilde{\pi}$ is realized.

B.1.1 (1) $\implies$ (2)

Since $\Delta(\Delta(\Omega))$ is a separable metric space under the weak* topology, Π is also a separable metric space. Thus, as (i) the von-Neumann Morgenstern axioms are satisfied, (ii) the upper- and lower-contour sets are closed, the standard conditions for the existence of a utility function (Debreu (1954), Theorem 2) are satisfied. The Blackwell ordering and (IR) properties are then inherited from Axioms (5) and (6), respectively. The same result of Debreu guarantees that this function will be unique up to affine transformations; if we set the cost of acquiring no information to be 0, then it is unique to linear transformations.

B.1.2 (2) $\implies$ (4)

Assign arbitrary values of $H(\mu_0)$ and $H(\delta_n)$ for $n < N$, where δ_n is the Dirac measure on state ω_n. For convenience, set $H(\mu_0) = 0$. Letting the fully informative experiment be π_∞,

[8] Recall that it is without loss of generality to consider $\mathcal{S} = \Delta(\Omega)$. It is important, though, that when we treat s as a signal, it only generates its corresponding value in $\Delta(\Omega)$ when the signal is generated with respect to μ_0, and not another prior, since it only generates that corresponding belief with respect to μ_0. Thus one must differentiate between its role as a signal and as a belief.

if the function H is to correctly give the costs, it must define $H(\delta_N)$ by

$$c(\pi_\infty) = \sum_{n=1}^{N} \mu_0(\omega_n) H(\delta_n)$$

Next, for any μ, consider the set of possible experiments that generate μ with positive probability, which I call $\{\pi_\mu\}$. Then the maximal possible probability that μ is generated is constrained by the state ω^* for which $\frac{\mu(\omega)}{\mu_0(\omega)}$ is the highest (by the law of total probability, there must be at least one for which this is greater than one, if $\mu \neq \mu_0$). By Bayes' rule, this maximal probability is $\frac{\mu_0(\omega^*)}{\mu(\omega^*)}$. Let π^*_μ be the experiment that generates μ with probability $\frac{\mu_0(\omega^*)}{\mu(\omega^*)}$, and otherwise fully reveals the state. Then $H(\mu)$ is defined by

$$c(\pi^*_\mu) = \frac{\mu_0(\omega^*)}{\mu(\omega^*)} H(\mu) + \sum_{i=1}^{N} H(\delta_i)\mu_0(\omega_i)[1 - \pi^*_\mu(s_\mu|\omega_i)]$$

$$\implies H(\mu) = \frac{\mu(\omega^*)}{\mu_0(\omega^*)}(c(\pi^*_\mu) - \sum_{i=1}^{N} H(\delta_i)\mu_0(\omega_i)[1 - \pi^*_\mu(s_\mu|\omega_i)])$$

where s_μ is the signal that yields posterior μ. Note that for $\omega_i = \omega^*$, $\pi^*_\mu(s_\mu|\omega) = 1$, so $\mu_0(\omega^*)[1 - \pi(s_\mu|\omega^*)]) = 0$.

I now show that any finite experiment π can be written as $E_{<\mu|\pi>}[H(\mu)]$. Consider some experiment π which generates a finite number of posteriors μ. I show that the cost of this experiment is pinned down by $c(\pi_\infty)$ and $c(\pi^*_\mu)$ for every μ in the support of τ_π.

Consider the following alternative experiment: given π, run π^*_μ with probability $\frac{\pi(s|\omega^*(s))}{\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s))}$ for each $s \in \mathcal{S}$, where ω^* is defined for each s as above. Note that by definition, $\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s)) > 1$. By (IR), this can be thought of as running some experiment $\hat{\pi}$ whose cost must be

$$c(\hat{\pi}) = \frac{\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s)) c(\pi^*_\mu)}{\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s))}$$

$$= \frac{1}{\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s))}[\sum_{\mu\in\text{supp}(\tau)} \tau_\pi(\mu) H(\mu) - \sum_i \mu_0(\omega_i) H(\delta_i)] + \sum_i \mu_0(\omega) H(\delta_i)$$

The same distribution over posteriors is also generated by running experiment π with probability $\frac{1}{\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s))}$, and experiment π_∞ with probability $1 - \frac{1}{\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s))}$. Therefore, by (IR),

$$\frac{1}{\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s))} c(\pi) + (1 - \frac{1}{\sum_{s\in\mathcal{S}} \pi(s|\omega^*(s))}) c(\pi_\infty) = c(\hat{\pi})$$

$$\implies c(\pi) = \sum_{\mu \in \text{supp}(\tau)} \tau_\pi(\mu) H(\mu)$$

To extend to the case of infinite experiments, one can approximate such experiments by discrete experiments π_k; since Π is a compact separable metric space, such approximations exist. Then by continuity,

$$c(\pi) = \lim_{k \to \infty} c(\pi_k)$$

$$= \lim_{k \to \infty} \int_{\Delta(\Omega)} H(\mu) d\tau_{\pi_k}(\mu)$$

$$= \int_{\Delta(\Omega}) H(\mu) d\tau_\pi(\mu)$$

$H(\mu_0)$ and $H(\delta_n)$ for $n < N$ were arbitrary, so there are N degrees of freedom in H. However, note that any $\tilde{H}(\mu) = H(\mu) + \sum_{n=1}^{N} \lambda_n \mu(\omega_n)$ results in the same cost, and has N degrees of freedom as well.[9] Moreover, a linear transformation of c by some constant λ_0 would be captured by multiplying H by λ_0. Thus H is essentially unique.

To see that H must be strictly convex, if not, then there exist two distributions μ, μ' such that for some $\alpha \in (0,1)$, $H(\alpha\mu + (1-\alpha)\mu') \geq \alpha H(\mu) + (1-\alpha)H(\mu')$. Then one could consider an experiment π that generates $\alpha\mu + (1-\alpha)\mu'$ with probability $\tau > 0$, and some experiment $\tilde{\pi}$ that is identical to π, except instead of generating $\alpha\mu + (1-\alpha)\mu'$, it generates μ with probability $\alpha\tau$, and μ' with probability $(1-\alpha)\tau$. $\tilde{\pi}$ is then Blackwell more informative that π, but the difference in costs is

$$c(\tilde{\pi}) - c(\pi) = \tau H(\alpha\mu + (1-\alpha)\mu') - \alpha\tau H(\mu) - (1-\alpha)\tau H(\mu') \leq 0$$

contradicting the fact that the cost is strictly increasing in the Blackwell order.

B.1.3 (4) $\implies$ (3)

Let $X = \Delta(\Omega)$. I show that there is some utility function u such that

$$\max_{x \in X} E_\mu[u(x, \omega)] = H(\mu) \tag{10}$$

Since H is strictly convex, there exists a subgradient at every point μ, which is unique almost everywhere (Rockafellar (1971), Theorem 25.5). Define this unique subgradient to be

[9] Note that $H(\mu_0)$ would be pinned down by the cost for experiment π_∞.

u_μ. Then the values of $u_\mu(\delta_n)$ for each ω_n are such that

$$\sum_{n=1}^{N} \mu(\omega) u_\mu(\omega) = H(\mu)$$

Moreover, since these are subgradients and H is strictly convex, each subgradient intersects H at exactly one point. Hence for a decision maker, the choice of subgradient u_μ is optimal if and only if one has the posterior μ, and so this gives us the decision problem which satisfies (9).

To see that this decision problem is essentially unique, first note that it is impossible to find a smaller space than $\Delta(\Omega)$, since one must have a distinct optimal action at each posterior $\mu \in \Delta(\omega)$ if H is to be strictly convex; otherwise, the action would also be optimal for convex combinations of the posterior, violating this strict convexity.

For any decision problem that satisfies (9), one can always relabel the optimal action at posterior μ as $x = \mu$, and so Criterion 3 (Irrelevance to Relabelling) of Section 2.3 is satisfied. One never needs more than one optimal action per posterior μ, so Criterion 2 (Irrelevance of Non-Optimal Actions) is also satisfied. Lastly, given that H is essentially unique, one can replace H with $\tilde{H}$, where $\tilde{H}(\mu) = \lambda_0 H(\mu) + \sum_{n=1}^{N-1} \lambda_n \mu(\omega_n)$ Then if one replaces u with $\tilde{u}$, where

$$\tilde{u}(x, \omega) = \lambda_0 u(x, \omega) + \lambda_n$$

yields

$$\max_{x \in X} E_\mu[\tilde{u}(x, \omega)] = \tilde{H}(\mu)$$

so Criterion 1 (Irrelevance to Affine Transformations) is satisfied.

B.1.4 (4) $\implies$ (5)

As seen in the proof that (2) $\implies$ (4), for every s, there exists some $\omega^*(s)$ such that $\frac{\mu(\omega^*(s)|s)}{\mu_0(\omega^*(s))}$ is maximal. Let $d\alpha(s) = d\pi(s|\omega^*(s))$. Given μ_0,

$$c(\pi) = \int_{\mathcal{S}} H(\mu(\cdot|s)) d\tau(\mu(\cdot|s)) - H(\mu_0)$$

As shown in the construction of H, it is without loss of generality to assume that $H(\mu_0) = 0$. Using Bayes' rule,

$$c(\pi) = \sum_{\omega \in \Omega} \int_{\mathcal{S}} H(\mu(\cdot|s)) \mu_0(\omega) d\pi(s|\omega)$$

Recall that μ is determined by the ratio $\frac{d\pi(s|\omega)}{d\pi(s|\omega')}$ for all ω, ω' (if π generates infinite signals, this is a Radon-Nikodym derivative).[10] Thus $\frac{d\pi(s|\omega)}{d\pi(s|\omega^*(s))}$ is well-defined and finite. This yields

$$c(\pi) = \sum_{\omega \in \Omega} \int_{\mathcal{S}} H(\mu(\cdot|s))\mu_0(\omega) \frac{d\pi(s|\omega)}{d\pi(s|\omega^*(s))} d\pi(s|\omega^*(s))$$

Let $\tilde{c}(\tilde{\pi}(s_\mu)) = \sum_{\omega \in \Omega} H(\mu)\mu_0(\omega) \frac{d\pi^*_\mu(s_\mu|\omega)}{d\pi^*_\mu(s_\mu|\omega^*(s_\mu))}$ Plugging in for $H(\mu)$ yields

$$\tilde{c}(\tilde{\pi}(s_\mu)) = \mu(\omega^*(s_\mu))(c(\pi^*_\mu) - \sum_{i=1}^{N} H(\delta_i)\mu_0(\omega_i)[1 - \pi^*_\mu(s_\mu|\omega_i)]) \sum_{\omega \in \Omega} \frac{\mu_0(\omega)}{\mu_0(\omega^*(s_\mu))} \frac{\pi^*_\mu(s_\mu|\omega)}{\pi^*_\mu(s_\mu|\omega^*(s_\mu))}$$

$$= (\sum_{\omega \in \Omega} \frac{\mu_0(\omega)}{\mu_0(\omega^*(s_\mu))} \frac{\pi^*_\mu(s_\mu|\omega)}{\pi^*_\mu(s_\mu|\omega^*(s_\mu))})^{-1} (c(\pi^*_\mu) - \sum_{i=1}^{N} H(\delta_i)\mu_0(\omega_i)[1 - \pi^*_\mu(s_\mu|\omega_i)]) \sum_{\omega \in \Omega} \frac{\mu_0(\omega)}{\mu_0(\omega^*(s_\mu))} \frac{\pi^*_\mu(s_\mu|\omega)}{\pi^*_\mu(s_\mu|\omega^*(s_\mu))}$$

$$= c(\pi^*_\mu) - \sum_{i=1}^{N} H(\delta_i)\mu_0(\omega_i)[1 - \pi^*_\mu(s_\mu|\omega_i)]$$

Recall that there still remain $N-1$ degrees of freedom in the choice of H. Thus, for each ω_i, let $H(\delta_i) = \frac{c(\pi_\infty)}{N\mu_0(\omega_i)}$; this satisfies the requisite condition that $\sum_{i=1}^{N} \mu_0(\omega_i)H(\delta_i) = c(\pi_\infty)$. Plugging this into $\tilde{c}$ yields

$$\tilde{c}(\tilde{\pi}(s)) = c(\pi^*_\mu) - \frac{c(\pi_\infty)}{N} \sum_{i=1}^{N} [1 - \pi^*_\mu(s_\mu|\omega_i)] \tag{11}$$

which is independent of the prior.

To demonstrate essential uniqueness, recall that there were N degrees of freedom in the choice of H. One of these degrees of freedom is taken by the fact that the uninformative signal must have cost 0. The other $N-1$ can be taken by the fact that any coefficients λ_i such that $\sum_{i=1}^{N} \lambda_i \mu_0(\omega_i) = 1$ can be used to set $H(\delta_i) = \lambda_i c(\pi_\infty)$, and will lead to a prior-independent representation.[11] If one scales c by a positive constant, one then scales $\tilde{c}$ by the same constant. $\square$

[10] As shown in Appendix A, this Radon-Nikodym derivative exists because the signal generates a valid posterior.

[11] Some of these possibilities lead to negative values of the cost $\tilde{c}$ for some realizations s (though when in combination with other signal realizations, any overall signal π will still lead to a positive cost). Note that with the specification of H as in the existence construction, $\tilde{c}$ will be positive, though.

B.2 Proof of Corollary 2

Assume throughout that π' is Blackwell more informative than π. By Theorem 1, for each DM, one can represent c by some posterior-separable measure of uncertainty H. Thus for any finite experiments π, π',

$$c_1(\pi') - c_1(\pi) \geq c_2(\pi') - c_2(\pi)$$

$$\implies \sum_{\text{supp}(\tau_{\pi'})} H_1(\mu)\tau_{\pi'}(\mu) - \sum_{\text{supp}(\tau_\pi)} H_1(\mu)\tau_\pi(\mu) \geq \sum_{\text{supp}(\tau_{\pi'})} H_2(\mu)\tau_{\pi'}(\mu) - \sum_{\text{supp}(\tau_\pi)} H_2(\mu)\tau_\pi(\mu) \tag{12}$$

Assume for now that inequality (12) is strict. Let $\tilde{H} = H_1 - H_2$. Then (12) reduces to

$$\sum_{\text{supp}(\tau_{\pi'})} \tilde{H}(\mu)\tau_{\pi'}(\mu) - \sum_{\text{supp}(\tau_\pi)} \tilde{H}(\mu)\tau_\pi(\mu) \geq 0$$

By Theorem 1, since $\tilde{H}$ is increasing in the Blackwell order, it is a convex posterior-separable measure of uncertainty.

To extend to the case where the inequality is weak, let

$$H_2^k(\mu) = H_2(\mu) + \frac{1}{2^k}\sum_{\omega\in\Omega} \mu(\omega)(1-\mu(\omega))$$

for $1 \leq k < \infty$. Then by the previous paragraph, $\tilde{H}^k$ (defined analogously) is a convex posterior-separable measure of uncertainty. In the limit, $\lim_{k\to\infty} \tilde{H}^k = \tilde{H}$ is a convex function.

Conversely, suppose that $\tilde{H} = H_1 - H_2$. Then if $\tilde{H}$ is convex, and $\tau_{\pi'}$ is a mean-preserving spread of τ_π, equation (11) follows. Letting $c_i(\pi) = \sum_{\text{supp}(\tau_\pi)} H_i(\mu)\tau_\pi(\mu) - H_i(\mu_0)$, this immediately implies that DM 1 has higher marginal costs of information acquisition.

B.3 Principal-agent model

Proof of Lemma 3: Suppose that it were optimal to generate $2 < N < \infty$ signals $\{s_0, ..., s_{N-1}\}$ that yield different posteriors given prior μ^*, where $\mu(e=1|s_k) > \mu(e=1|s_n)$ whenever $k > n$. By standard arguments, one can assume that $t(s_0) = 0$, and so

$$t(s_1) = u_A^{-1}\left(\frac{c_A + u_A(0)(\pi(s_0|e=0) - \pi(s_0|e=1)}{\pi(s_1|e=1) - \pi(s_1|e=0)}\right)$$

In order for effort to be both optimal for the principal, and incentive-compatible for the agent, one must have that for any $k > n$, $t(s_k) > t(s_n)$. Otherwise, there exists the

following improvement: suppose that $\pi(s_k|e=1) \geq \pi(s_n|e=1)$. Then the principal can offer the following alternative compensation scheme: if s_n is observed, then offer $t(s_k)$; if s_k is observed, then with probability $\frac{\pi(s_n|e=1)}{\pi(s_k|e=1)}$, offer $t(s_n)$, and otherwise offer $t(s_k)$. This preserves the expected utility for the agent conditional on exerting effort. However, if the agent does not exert effort, then since $\frac{\pi(s_k|e=0}{\pi(s_n|e=0)} < \frac{\pi(s_k|e=0)}{\pi(s_n|e=0)}$, the expected utility from $e=0$ decreases. This relaxes the IC constraint, and so the principal can pay the agent less. This therefore would represent an improvement to the principal. The argument is analogous if $\pi(s_k|e=1) \geq \pi(s_n|e=1)$, with the only difference being that one offers $t(s_n)$ if s_k is observed, and randomizes if s_n is observed.

Now consider, instead, the alternative signal structure in which the lowest two signals, s_0, s_1, are replaced by one signal, $\tilde{s}_0$, and otherwise $\mathcal{S}$ remains unchanged. Meanwhile, set $t(\tilde{s}_0) = 0$, and increase $t(s_{N-1})$ by $\frac{\pi(s_{N-1}|e=1)}{\pi(s_1|e=1)}t(s_1)$. By Bayes' rule, $\frac{\pi(s_N|e=0)}{\pi(s_1|e=0)} < \frac{\pi(s_N|e=1)}{\pi(s_1|e=1)}$, and so $\pi(\tilde{s}_0|e=0) > \pi(\tilde{s}_0|e=1)$. Therefore, the expected payoff to the agent from choosing $e=1$ is unchanged, while it decreases for choosing $e=0$. Thus this modification of the monitoring structure would be a (weak) improvement, and so it is without loss of generality to have binary signals.

To extend this observation to rule out the need for infinite signals, suppose that $(\pi, \mathcal{S})$ (where $\mathcal{S}$ be set to be $[0,1]$ without loss of generality, i.e. $s = \mu(\cdot|s)$) is optimal. Let $\{\psi_k\}_{k=1}^{\infty}$ be a sequence of simple functions that approximate $t(\cdot)$. More precisely, let

$$\psi_k(s) = \max_{0 \leq i \leq k^2} \{\frac{i}{k} \leq t(s)\}$$

By the dominated convergence theorem,

$$\lim_{k\to\infty} \int_0^1 \psi_k(s) d\tau(s) = \int_0^1 t(s) d\tau(s)$$

$$\lim_{k\to\infty} \int_0^1 \psi_k(s) d\pi(s|e) = \int_0^1 t(s) d\pi(s|e)$$

By continuity of u, this also implies that

$$\lim_{k\to\infty} \int_0^1 u(\psi_k(s)) d\pi(s|e) = \int_0^1 u(t(s)) d\pi(s|e) \tag{13}$$

Note that the approximation by simple functions is equivalent to a signal structure $(\pi_k, \mathcal{S}_k)$ with $|\mathcal{S}| \leq k^2$: it would not affect anyone's payoff to merge all signals which generate the same value of ψ_k. Thus $c(\pi_k) \leq c(\pi)$.

Effort may not be incentive compatible under $(\pi_k, \mathcal{S}_k)$. If not, then to rectify this, define

Δ_k implicitly so as to solve

$$\sum_{s_k \in \mathcal{S}_k} u(t(s_k))(\pi_k(s_k|0) - \pi_k(s_k|1)) = \sum_{s_k \in \mathcal{S}_k : \pi_k(s_k|1) > \pi_k(s_k|0)} u(t_k(s_k) + \Delta_k)(\pi_k(s_k|1) - \pi_k(s_k|0))$$

Note that by the law of total probability, $\pi_k(s_k|1) > \pi_k(s_k|0)$ for some s_k. By equation (13) and the continuity of u, it follows that $\lim_{k\to\infty} \Delta_k = 0$.

Let $(t_2^*, \pi_2^*, \mathcal{S}_2^*)$ be the optimal binary policy for the principal. As argued above, this policy is better for the principal than any other finite policy. Thus,

$$1 - c(\pi^*) - \sum_{s \in \mathcal{S}_2^*} t_2^*(s)\pi^*(s|1) \geq \liminf_{k\to\infty} [\sum_{s \in \mathcal{S}_k} \psi_k(s)\pi_k(s|1) - c(\pi_k)]$$

$$= \liminf_{k\to\infty} [\sum_{s_k \in \mathcal{S}_k : \pi_k(s_k|1) \leq \pi_k(s_k|0)} \psi_k(s)\pi_k(s|1) + \sum_{s_k \in \mathcal{S}_k : \pi_k(s_k|1) > \pi_k(s_k|0)} (\psi_k(s) + \Delta_k)\pi_k(s|1) - c(\pi_k)]$$

$$\geq \int t(s) d\pi(s|1) - c(\pi) \, \square$$

B.4 Proof of Proposition 4:

It suffices to show that this is not possible for $N = 2$. Suppose that such a representation exists. This implies that the (IR) property holds, and so I construct an experiment that can be represented by a randomization in two different ways, yet yields different costs as a result. For shorthand, denote $\mu_0 \equiv \mu_0(\omega_2)$. Define signal $(\pi, \{s_1, s_2\})$ such that

$$\pi(s_2|\omega) = \begin{cases} p, & \omega = \omega_2 \\ 1 - p, & \omega = \omega_1 \end{cases}$$

and vice versa for $s = s_1$. Then if we consider some signal π^2 which is equivalent to performing π twice (i.e. π is run again at any posterior μ generated by π), the posterior conditional on getting s_2 twice will be $\frac{\mu_0 p^2}{\mu_0 p^2 + (1-\mu_0)(1-p)^2}$; that from getting s_1 twice, $\frac{\mu_0 (1-p)^2}{(1-\mu_0)p^2 + \mu_0(1-p)^2}$; and otherwise, μ_0. These occur with respective probabilities $\mu_0 p^2 + (1-\mu_0)(1-p)^2$, $(1-\mu_0)p^2 + \mu_0(1-p)^2$, and $2p(1-p)$. We can also express $\pi^2 = 2p(1-p)\pi_0 + (1 - 2p + 2p^2)\tilde{\pi}^2$, where

$$\tilde{\pi}^2(s_2|\omega) = \begin{cases} \frac{p^2}{1-2p+2p^2}, & \omega = \omega_2 \\ \frac{(1-p)^2}{1-2p+2p^2}, & \omega = \omega_1 \end{cases}$$

By the hypothesis of the theorem, $2c(\pi) = c(\pi^2)$. So, by (IR), we get

$$2c(\pi) = c(\pi^2) = (1 - 2p + 2p^2)c(\tilde{\pi}^2)$$

Hence

$$c(\tilde{\pi}^2) = \frac{2}{1 - 2p + 2p^2}c(\pi)$$

More generally, define the experiment $(\tilde{\pi}^k, \{s_1, s_2\})$, where

$$\tilde{\pi}^k(s_2, \omega) = \begin{cases} \frac{p^k}{p^k+(1-p)^k}, & \omega = \omega_2 \\ \frac{(1-p)^k}{p^k+(1-p)^k}, & \omega = \omega_1 \end{cases}$$

Next, define experiment $(\pi^k, \{s_1, s_2\})$ inductively by looking at the outcomes when first experiment $\tilde{\pi}^{k-1}$ and then π are performed. The posteriors are then distributed according to

$$\tau_{\pi^k}(\mu) = \begin{cases} \frac{\mu_0 p^k}{\mu_0 p^k+(1-\mu_0)(1-p)^k}, & \frac{p^k\mu_0+(1-\mu_0)(1-p)^k}{p^{k-1}+(1-p)^{k-1}} \\ \frac{\mu_0 p^{k-2}}{\mu_0 p^{k-2}+(1-\mu_0)(1-p)^{k-2}}, & \frac{p^{k-1}(1-p)\mu_0+(1-\mu_0)(1-p)^{k-1}p}{p^{k-1}+(1-p)^{k-1}} \\ \frac{\mu_0(1-p)^k}{(1-\mu_0)p^k+\mu_0(1-p)^k}, & \frac{p^k(1-\mu_0)+\mu_0(1-p)^k}{p^{k-1}+(1-p)^{k-1}} \\ \frac{\mu_0(1-p)^{k-2}}{\mu_0(1-p)^{k-2}+(1-\mu_0)p^{k-2}}, & \frac{p^{k-1}(1-p)(1-\mu_0)+\mu_0(1-p)^{k-1}p}{p^{k-1}+(1-p)^{k-1}} \end{cases}$$

Putting this all together, by (IR), we have

$$c(\tilde{\pi}^{k-1}) + c(\pi) = \frac{(p^k + (1-p)^k)}{p^{k-1} + (1-p)^{k-1}}c(\tilde{\pi}^k) + (1 - \frac{p^k + (1-p)^k}{p^{k-1} + (1-p)^{k-1}})c(\tilde{\pi}^{k-2}) \qquad (14)$$

For $k = 3$, (14) states

$$c(\tilde{\pi}^2) + c(\pi) = \frac{p^3 + (1-p)^3}{p^2 + (1-p)^2}c(\tilde{\pi}^3) + \frac{p^2 + (1-p)^2 - p^3 - (1-p)^3}{p^2 + (1-p)^2}c(\pi)$$

$$2c(\pi) + (p^2 + (1-p)^2)c(\pi) = (p^3 + (1-p)^3)c(\tilde{\pi}^3) + (p^2 + (1-p)^2 - p^3 - (1-p)^3)c(\pi)$$

$$\implies c(\tilde{\pi}^3) = (1 + \frac{2}{p^3 + (1-p)^3})c(\pi)$$

For $k = 4$, (14) states

$$c(\tilde{\pi}^3) + c(\pi) = \frac{p^4 + (1-p)^4}{p^3 + (1-p)^3}c(\tilde{\pi}^4) + \frac{p^3 + (1-p)^3 - p^4 - (1-p)^4}{p^3 + (1-p)^3}c(\tilde{\pi}^2)$$

$$c(\tilde{\pi}^4) = \frac{(p^3 + (1-p)^3 + 2 + p^3 + (1-p)^3)(p^2 + (1-p)^2) - 2p^3 - 2(1-p)^3 + 2p^4 + 2(1-p)^4}{p^4 + (1-p)^4}c(\pi)$$

$$= \frac{c(\pi)}{p^4 + (1-p)^4}[(4 - 6p + 6p^2)(1 - 2p + 2p^2) - 2p + 6p^2 - 8p^3 + 4p^4]$$

$$= \frac{c(\pi)}{p^4 + (1-p)^4}[4 - 16p + 32p^2 - 32p^3 + 16p^4] \tag{15}$$

At the same time, by (IR), $c(\tilde{\pi}^4)$ has to be consistent with running experiment $\tilde{\pi}^2$ twice. Thus

$$c(\tilde{\pi}^4) = \frac{2}{1 - 2\frac{p^2}{1-2p+2p^2} + 2(\frac{p^2}{1-2p+2p^2})^2} c(\tilde{\pi}^2)$$

$$= \frac{4}{p^2 + (1-p)^2 - 2p^2 + \frac{2p^4}{p^2+(1-p)^2}} c(\pi)$$

$$= \frac{4p^2 + 4(1-p)^2}{(1-2p)(p^2 + (1-p)^2) + 2p^4} c(\pi)$$

$$= \frac{4p^2 + 4(1-p)^2}{p^4 + (1-p)^4} c(\pi) \tag{16}$$

But equation (15) is inconsistent with (16), and so there cannot be such $c(\cdot)$.