The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Information Acquisition Under Resource Limitations in a Noisy Environment

Matvey Soloviev
Computer Science Department
Cornell University
msoloviev@cs.cornell.edu

Joseph Y. Halpern
Computer Science Department
Cornell University
halpern@cs.cornell.edu

Abstract

We introduce a theoretical model of information acquisition under resource limitations in a noisy environment. An agent must guess the truth value of a given Boolean formula φ after performing a bounded number of noisy tests of the truth values of variables in the formula. We observe that, in general, the problem of finding an optimal testing strategy for φ is hard, but we suggest a useful heuristic. The techniques we use also give insight into two apparently unrelated, but well-studied problems: (1) *rational inattention* (the optimal strategy may involve hardly ever testing variables that are clearly relevant to φ) and (2) what makes a formula hard to learn/remember.

1 Introduction

Decision-making is typically subject to resource constraints. However, an agent may be able to choose how to allocate his resources. We consider a simple decision-theoretic framework in which to examine this resource-allocation problem. To motivate the framework, consider an animal that must decide whether some food is safe to eat. We assume that "safe" is characterised by a known Boolean formula φ, which depends on pertinent variables such as presence of unusual smells or signs of other animals consuming the same food. The animal can perform a limited number of tests of the variables in φ, but these tests are noisy; if a test says that a variable v is true, that does not mean that v is true, but only that it is true with some probability. After the agent has exhausted his test budget, he must either guess the truth value of φ or choose not to guess. Depending on his choice, he gets a payoff. In this example, guessing that φ is true amounts to guessing that the food is safe to eat. There will be a small positive payoff for guessing "true" if the food is indeed safe, but a large negative payoff for guessing "true" if the food is not safe to eat. In this example we can assume a payoff of 0 if the agent guesses "false" or does not guess, since both choices amount to not eating the food.

We are interested in optimal strategies for this decision; that is, what tests should the agent perform and in what order. Unfortunately (and perhaps not surprisingly), as we show, finding an optimal strategy (i.e., one that obtains the highest expected payoff) is infeasibly hard. We provide a heuristic that guarantees a positive expected payoff whenever the optimal strategy gets a positive expected payoff. Our analysis of this strategy also gives us the tools to examine two other problems of interest.

The first is *rational inattention*, the notion that in the face of limited resources it is sometimes rational to ignore certain sources of information completely. There has been a great deal of interest recently in this topic in economics (Sims 2003; Wiederholt 2010). Here we show that optimal testing strategies in our framework exhibit what can reasonably be called rational inattention (which we typically denote RI from now on). Specifically, our experiments show that for a substantial fraction of formulae, an optimal strategy will hardly ever test variables that are clearly relevant to the outcome. (Roughy speaking, "hardly ever" means that as the total number of tests goes to infinity, the fraction of tests devoted to these relevant variables goes to 0.) For example, consider the formula $v_1 \vee v_2$. Suppose that the tests for v_1 and v_2 are equally noisy, so there is no reason to prefer one to the other for the first test. But for appropriate choices of payoffs, we show that if we start by testing v_2, then all subsequent tests should also test v_2 as long as v_2 is observed to be true (and similarly for v_1). Thus, with positive probability, the optimal strategy either ignores v_1 or ignores v_2. Our formal analysis allows us to conclude that this is a widespread phenomenon.

The second problem we consider is what makes a concept (which we can think of as being characterised by a formula) hard. To address this, we use our framework to define a notion of hardness. We show that, according to this definition, XORs (i.e., formulae of the form $v_1 \oplus \cdots \oplus v_n$, which are true exactly if an odd number of the v_i's are true) and their negations are the hardest formulae. We compare this notion to other notions of hardness of concepts considered in the cognitive psychology literature (e.g., (Feldman 2006; Love, Medin, and Gureckis 2004; Shepard, Hovland, and Jenkins 1961)).

2 Information-acquisition games

We model the *information-acquisition game* as a single-player game against nature. The game is characterised by five parameters:

- a Boolean formula φ that mentions variables $v_1, \ldots, v_n$ for some $n > 0$;

- a probability distribution D on truth assignments to $\{v_1, \ldots, v_n\}$;
- a bound k on the number of tests;
- an *accuracy vector* $\vec{\alpha} = (\alpha_1, \ldots, \alpha_n)$, with $0 \leq \alpha_i \leq 1/2$ (explained below);
- payoffs (g, b), where $g > 0 > b$ (also explained below).

We denote this game as $G(\varphi, D, k, \vec{\alpha}, g, b)$.

In the game $G(\varphi, D, k, \vec{\alpha}, g, b)$, nature first chooses a truth assignment to the variables $v_1, \ldots, v_n$ according to distribution D. While the parameters of the game are known to the agent, the assignment chosen by nature is not. For the next k rounds, the agent then chooses one of the n variables to test (perhaps as a function of history), and nature responds with either T or F. The agent then must either guess the truth value of φ or choose not to guess. We view a truth assignment as a function from variables to truth values ($\{T, F\}$); we can also view the formula φ itself as a function from truth assignments to truth values. If the agent chooses to test v_i, then nature returns $A(v_i)$ (the right answer) with probability $1/2 + \alpha_i$ (and thus returns $\neg A(v_i)$ with probability $1/2 - \alpha_i$).[1] Finally, if the agent choses not to guess at the end of the game, his payoff is 0. If he chooses to guess, then his payoff is g (good) if he guesses $\varphi(A)$ (i.e., if he guesses the truth value of φ correctly, given that A is the actual truth assignment) and b (bad) if he guesses $\neg\varphi(A)$. A strategy for an agent in this game is just a function that determines which test the agent performs after observing each test-outcome sequence of length $< k$, together with a final action for each test-outcomes sequence of length k.

Example 2.1. Consider the information-acquisition game over the formula $v_1 \vee v_2$, with $k = 2$ tests, a uniform distribution on truth assignments, accuracy vector $(1/4, 1/4)$, correct-guess reward $g = 1$ and wrong-guess penalty $b = -16$. As we show (see Appendix A in the full paper) this game has two optimal strategies:

1. test v_1 twice, guess T if both tests came out T, and make no guess otherwise;
2. test v_2 twice, guess T if both tests came out T, and make no guess otherwise.

□

Thus, in this game, an optimal strategy either ignores v_1 or ignores v_2. As we show in the full paper, the strategy "test v_1 and then v_2, guess T if both tests came out T" is strictly worse than these two; in fact, its expected payoff is negative!

If we increase k, the situation becomes more nuanced. For instance, if $k = 4$, an optimal strategy tests v_1 once, and if the test comes out F, tests v_2 three times and guesses T if all three tests came out T. However, it always remains optimal to keep testing one variable as long as the tests keep coming out true. That is, all optimal strategies exhibit RI in the sense that there are test outcomes that result in either v_1 never being tested or v_2 never being tested, despite their obvious relevance to $v_1 \vee v_2$.

We will frequently talk about the probability of various events over the course of a run of the game, and many of the probabilities we care about depend only on a few parameters of the game. Formally, we embed the traces of all information-acquisition games on formulae in n variables in a common sample space Ω_n, whose elements are tuples of the form

$$(A, v_{i_1} \approx b_1, \ldots, v_{i_k} \approx b_k, a),$$

where A is the assignment of truth values to the n variables chosen by nature, $v_{i_j} \approx b_j$ indicates that the jth test was performed on variable v_{i_j} and that nature responded with the test outcome b_j, and a is the final agent action of either making no guess or guessing some truth value for the formula. A game $G(\varphi, D, k, \vec{\alpha}, g, b)$ and agent strategy σ then induce a probability $\Pr_{G,\sigma}$ on this sample space. The only features of the game G that affect the probability are the prior distribution D and the accuracy vector α, so we write $\Pr_{D,\alpha,\sigma}(\varphi)$ rather than $\Pr_{G,\sigma}(\varphi)$. If some component of the subscript does not affect the probability, then we typically omit it. In particular, we show in Appendix B of the full paper that the strategy σ does not affect $\Pr_{G,\sigma}(\varphi|S)$, so we write $\Pr_{D,\vec{\alpha}}(\varphi|S)$. Finally, the utility (payoff) received by the agent at the end of the game is a real-valued random variable that depends on parameters b and g. We can define the expected utility $\mathbb{E}_{G,\sigma}(\text{payoff})$ as the expectation of this random variable.

3 Determining optimal strategies

It is straightforward to see that the game tree for the game $G(\varphi, D, k, \vec{\alpha}, g, b)$ has $3(2^n)(2n)^k$ leaves: there is a branching factor of 2^n at the root (since there are 2^n truth assignments) followed by k branching factors of n (for the n variables that the agent can choose to test) and 2 (for the two possible outcomes of a test). At the end there are three choices (don't guess, guess T, and guess F). A straightforward backward induction can then be used to compute the optimal strategy. Unfortunately, the complexity of this approach is polynomial in the number of leaves; it quickly grows infeasible as k grows.

In general, it is unlikely that the dependency on 2^n can be removed. In the special case that $b = -\infty$ and $\alpha_i = \frac{1}{2}$ for all i (so tests are perfectly accurate, but the truth value of the formula must be established for sure), determining whether a strategy of length k gets a positive expected payoff reduces to the problem of finding a conjunction of length k that implies a given Boolean formula. Umans (1999) showed that this problem is Σ_2^p-complete, that is, lies in a complexity class that is at least as hard as both NP and co-NP.

A simple heuristic (that is independent of φ) would simply be to test each variable in φ k/n times, and then choose the action that maximises the expected payoff given the observed test outcomes. We can calculate in time polynomial in k and n the expected payoff of a guess, conditional on a sequence of test outcomes. Since determining the best guess involves checking the likelihood of each of the 2^n truth assignments conditional on the outcomes, this approach takes time polynomial in k and 2^n. We are most interested in formulae where n

[1] Note that this means that the probability of a false positive and that of a false negative are the same. While we could easily extend the framework so as to allow the accuracy in a test on a variable v to depend on whether $A(v)$ is T or F, doing so would complicate notation and distract from the main points that we want to make.

is small, so this time complexity would be acceptable. However, this approach can be arbitrarily worse than the optimum. As we observed in Example 2.1, the expected payoff of this strategy is negative, while there is a strategy that has positive expected payoff.

An arguably somewhat better heuristic, which we call the *random-test heuristic*, is to choose, at every step, the next variable to test uniformly at random, and again, after k observations, choosing the action that maximises the expected payoff. This heuristic clearly has the same time complexity as the preceding one, while working better in information-acquisition games that require an unbalanced approach to testing.

Proposition 3.1. *If there exists a strategy that has positive expected payoff in the information-acquisition game G, then the random-test heuristic has positive expected payoff.*

To prove Proposition 3.1, we need a preliminary lemma. Intuitively, an optimal strategy should try to generate test-outcome sequences S that maximise $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2|$, since the larger $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2|$ is, the more certain the agent is regarding whether φ is true or false. The following lemma characterises how large $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2|$ has to be to get a positive expected payoff.

Definition 3.2. Let $q(b,g) = \frac{b+g}{2(b-g)}$ be the *threshold* associated with payoffs b, g. □

Lemma 3.3. *The expected payoff of $G(\varphi, D, k, \vec{\alpha}, g, b)$ when making a guess after observing a sequence S of test outcomes is positive if and only if*

$$|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2| > q(b,g). \tag{1}$$

Proof. The expected payoff when guessing that the formula is true is

$$g \cdot \Pr_{D,\vec{\alpha}}(\varphi \mid S) + b \cdot (1 - \Pr_{D,\vec{\alpha}}(\varphi \mid S)).$$

This is greater than zero iff

$$(g-b)\Pr_{D,\vec{\alpha}}(\varphi \mid S) + b > 0,$$

that is, iff

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2 > \frac{b}{b-g} - \frac{1}{2} = q(b,g).$$

When guessing that the formula is false, we simply exchange $\Pr_{D,\vec{\alpha}}(\varphi \mid S)$ and $1 - \Pr_{D,\vec{\alpha}}(\varphi \mid S)$ in the derivation. So the payoff is then positive iff

$$(1 - \Pr_{D,\vec{\alpha}}(\varphi \mid S)) - \frac{1}{2} = -(\Pr_{D,\vec{\alpha}}(\varphi \mid S) - \frac{1}{2}) > q(b,g).$$

Since $|x| = \max\{x, -x\}$, at least one of these two inequalities must hold if (1) does, so the corresponding guess will have positive expected payoff. Conversely, since $|x| \geq x$, either inequality holding implies (1). □

Proof of Proposition 3.1. Suppose that σ is a strategy for G with positive expected payoff. The outcome sequences of length k partition the space of paths in the game tree, so we have

$$\sum_{\{S:|S|=k\}} \Pr_{D,\vec{\alpha},\sigma}(S)\mathbb{E}_{G,\sigma}(\text{payoff} \mid S).$$

Since the payoff is positive, at least one of the summands on the right must be, say the one due to the sequence S^*. By Lemma 3.3, $|\Pr_{D,\vec{\alpha}}(\varphi \text{ is true} \mid S^*) - 1/2| > q(b,g)$.

Let τ denote the random-test heuristic. Since τ chooses the optimal action after making k observations, it will not get a negative expected payoff for any sequence S of k test outcomes (since it can always obtain a payoff of 0 by choosing not to guess). On the other hand, with positive probability, the variables that make up the sequence S^* will be chosen and the outcomes in S^* will be observed for these tests; that is $\Pr_{D,\vec{\alpha},\tau}(S^*) > 0$. It follows from Lemma 3.3 that $\mathbb{E}_{G,\tau}(\text{payoff} \mid S^*) > 0$. Thus, $\mathbb{E}_{G,\tau}(\text{payoff}) > 0$, as desired. □

4 Rational inattention

We might think that an optimal strategy for learning about φ would test all variables that are relevant to φ (given a sufficiently large test budget). As shown in Example 2.1, this may not be true. For example, an optimal k-step strategy for $v_1 \vee v_2$ can end up never testing v_1, no matter what the value of k, if it starts by testing v_2 and keeps discovering that v_2 is true. It turns out that RI is quite widespread.

It certainly is not surprising that if a variable v does not occur in φ, then an optimal strategy would not test v. More generally, it would not be surprising that a variable that is not particularly relevant to φ is not tested too often, perhaps because it makes a difference only in rare edge cases. In the foraging animal example from the introduction, the possibility of a human experimenter having prepared a safe food to look like a known poisonous plant would impact whether it is safe to eat, but is unlikely to play a significant role in day-to-day foraging strategies. What might seem more surprising is if a variable v is (largely) ignored while another variable v' that is no more relevant than v is tested. This is what happens in Example 2.1; although we have not yet defined a notion of relevance, symmetry considerations dictate that v_1 and v_2 are equally relevant to $v_1 \vee v_2$, yet an optimal strategy might ignore one of them.

The phenomenon of rational inattention observed in Example 2.1 is surprisingly widespread. To make this claim precise, we need to define "relevance". There are a number of reasonable ways of defining it; we focus on one below, although our results hold for other reasonable definitions too. The definition of the relevance of v to φ that we use counts the number of truth assignments for which changing the truth value of v changes the truth value of φ.

Definition 4.1. Define the relevance ordering $\leq_\varphi$ on the variables in φ by taking

$$\begin{aligned} & v \leq_\varphi v' \text{ iff} \\ & |\{A : \varphi(A[v \mapsto \mathrm{T}]) \neq \varphi(A[v \mapsto \mathrm{F}])\}| \\ \leq\ & |\{A : \varphi(A[v' \mapsto \mathrm{T}]) \neq \varphi(A[v' \mapsto \mathrm{F}])\}|, \end{aligned}$$

where $A[v \mapsto b]$ is the assignment that agrees with A except that it assigns truth value b to v.

Thus, rather than saying that v is or is not relevant to φ, we can say that v is (or is not) at least as relevant to φ as v'. Considering the impact of a change in a single variable to

the truth value of the whole formula in this fashion has been done both in the cognitive science and the computer science literature: for example, Vigo (2011) uses the *discrete (partial) derivative* to capture this effect, and Lang et al. (2003) define the related notion of *Var-independence*.

We could also consider taking the probability of the set of truth assignments where a variable's value makes a difference, rather than just counting how many such truth assignments there are. This would give a more detailed quantitative view of relevance, and is essentially how relevance is considered in Bayesian networks. Irrelevance is typically identified with independence. Thus, v is relevant to φ if a change to v changes the probability of φ. (See Druzdzel and Suermondt (1994) for a review of work on relevance in the context of Bayesian networks.) We did not consider a probabilistic notion of relevance because then the relevance order would depend on the game (specifically, the distribution D, which is one of the parameters of the game). Our definition makes the relevance order depend only on φ. That said, we believe that essentially the same results as those we prove would obtain for a probabilistic notion of relevance ordering.

Roughly speaking, φ exhibits RI if, for all optimal strategies σ for the game $G(\varphi, D, k, \vec{\alpha}, b, g)$, with probability 1, σ tests a variable v' frequently while hardly ever testing a variable v that is at least as relevant to φ as v'. We still have to make precise "hardly ever", and explain how the claim depends on the choice of D, $\vec{\alpha}$, k, b, and g. For the latter point, note that in Example 2.1, we had to choose b and g appropriately to get RI. This turns out to be true in general; given D, k, and $\vec{\alpha}$, the claim holds only for an appropriate choice of b and g that depends on these. In particular, for any fixed choice of b and g that depends only on k and $\vec{\alpha}$, there exist choices of priors D for which the set of optimal strategies is fundamentally uninteresting: we can simply set D to assign a probability to some truth assignment A that is so high that no k test outcomes could make it rational to make any other guess than that the truth value of the formula is $\varphi(A)$.

Another way that the set of optimal strategies can be rendered uninteresting is when, from the outset, there is no hope of obtaining sufficient certainty of the formula's truth value with the k tests available. Similarly to when the truth value is a foregone conclusion, in this situation, an optimal strategy can perform arbitrary tests, as long as it makes no guess at the end. More generally, even when in general the choice of variables to test does matter, a strategy can reach a situation where there is sufficient uncertainty that no future test outcome could affect the final choice. Thus, a meaningful definition of RI that is based on the variables tested by optimal strategies must consider only tests performed in those cases in which a guess actually should be made (because the expected payoff of the optimal strategy is positive).[2] We now make these ideas precise.

[2]One way to avoid these additional requirements is to modify the game so that performing a test is associated has a small but positive cost, so that an optimal strategy avoids frivolous testing when the conclusion is foregone. The definitions we use have essentially the same effect, and are easier to work with.

Definition 4.2. A function $f : \mathbb{N} \to \mathbb{N}$ is *negligible* if $\lim_{k\to\infty} f(k)/k = 0$. □

The idea is that φ exhibits RI if, as the number k of tests allowed increases, the fraction of times that some variable v is tested is negligible relative to the number of times that another variable v' is tested, although v is at least as relevant to φ as v'. We actually require slightly more: we want v' to be tested a linear number of times (i.e., at least ck times, for some constant $c > 0$).

Since we do not want our results to depend on correlations between variables, we restrict attention to probability distributions D on truth assignments that are product distributions.

Definition 4.3. A probability distribution D on truth assignments to n variables is a *product distribution* if there exist probability distributions D_i on truth assignments to v_i for $i = 1, \ldots, n$ such that $D = D_1 \times \cdots \times D_n$. □

As discussed earlier, to get an interesting notion of RI, we need to allow the choice of payoffs b and g to depend on the prior distribution D; for fixed b, g, and testing bound k, if the distribution D places sufficiently high probability on a single assignment, no k outcomes can change the agent's mind. For similar reasons, we do not want to allow D to give a single assignment probability 1. More generally, assigning prior probability 1 to any one variable being true or false means that no tests will change the agent's mind about that variable, and so testing it is pointless (and the game is therefore equivalent to one played on the formula in $n - 1$ variables where this variable has been replaced by the appropriate truth value). We say that a probability distribution that gives all truth assignments positive probability is *open-minded*.

With all these considerations in hand, we can finally define RI formally.

Definition 4.4. The formula φ *exhibits rational inattention* if, for all open-minded product distributions D and uniform accuracy vectors $\vec{\alpha}$ (those with $(\alpha_1 = \ldots = \alpha_n)$), there exists a negligible function f and a constant $c > 0$ such that for all k, there are payoffs b and g such that all optimal strategies in the information-acquisition game $G(\varphi, D, k, \vec{\alpha}, b, g)$ have positive expected payoff and, in all runs of the game, depending on outcomes of tests, either make no guess or

- test a variable v' at least ck times, but
- test a variable v such that $v' \leq_\varphi v$ at most $f(k)$ times.

□

We can check in a straightforward way whether some natural classes of formulae exhibit RI in the sense of this definition.

Example 4.5. (Rational inattention)

1. Conjunctions $\varphi = \bigwedge_{i=1}^{N} \ell_i$ and disjunctions $\varphi = \bigvee_{i=1}^{N} \ell_i$ of $N \geq 2$ literals (variables $\ell_i = v_i$ or their negations $\neg v_i$) exhibit RI. In each case, we can pick b and g such that all optimal strategies pick one variable and focus on it, either to establish that the formula is true (for disjunctions) or that it is false (for conjunctions). By symmetry, all variables v_i and v_j are equally relevant, so $v_i \leq_\varphi v_j$.

2. The formulae v_i and $\neg v_i$ do not exhibit RI. There is no variable $v \neq v_i$ such that $v_i \leq_{(\neg)v_i} v$, and for all choices of b and g, the strategy of testing only v_i and ignoring all other variables (making an appropriate guess in the end) is clearly optimal for $(\neg)v_i$.
3. More generally, we can say that all XORs in ≥ 0 variables do not exhibit RI. For the constant formulae T and F, any testing strategy that "guesses" correctly is optimal; for any XOR in more than one variable, an optimal strategy must test all of them as any remaining uncertainty about the truth value of some variable leads to at least equally great uncertainty about the truth value of the whole formula. Similarly, negations of XORs do not exhibit RI. Together with the preceding two points, this means that the only formulae in 2 variables exhibiting rational inattention are the four conjunctions $\ell_1 \wedge \ell_2$ and the four disjunctions $\ell_1 \vee \ell_2$ in which each variable occurs exactly once and may or may not be negated.
4. For $n > 2$, formulae φ of the form $v_1 \vee (\neg v_1 \wedge v_2 \wedge \ldots \wedge v_n))$ do not exhibit RI. Optimal strategies that can attain a positive payoff at all will start by testing v_1; if the tests come out true, it will be optimal to continue testing v_1, ignoring $v_2 \ldots v_n$. However, for formulae φ of this form, v_1 is strictly more relevant than the other variables: there are only 2 assignments where changing v_i flips the truth value of the formula for $i > 1$ (the two where $v_1 \mapsto F$ and $v_j \mapsto T$ for $j \notin \{1, i\}$) but $2^n - 2$ assignments where changing v_1 does (all but the two where $v_j \mapsto T$ for $j \neq 1$). Hence, in the event that all these tests actually succeed, the only variables that are ignored are not at least as relevant as the only one that isn't, so φ does not exhibit RI.

□

Unfortunately, as far as we know, determining the optimal strategies is hard in general. To be able to reason about whether φ exhibits RI in a tractable way, we find it useful to consider optimal test-outcome sequences.

Definition 4.6. A sequence S of test outcomes is *optimal* for a formula φ, prior D, and accuracy vector $\vec{\alpha}$ if it minimises the conditional uncertainty about the truth value of φ among all test-outcome sequences of the same length. That is,

$$\left|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - \frac{1}{2}\right| \geq \left|\Pr_{D,\vec{\alpha}}(\varphi \mid S') - \frac{1}{2}\right|$$

for all S' with $|S'| = |S|$.

Using this definition, we can derive a sufficient (but not necessary!) condition for formulae to exhibit RI.

Proposition 4.7. *Suppose that, for a given formula φ, for all open-minded product distributions D and uniform accuracy vectors $\vec{\alpha}$, there exists a negligible function f and a constant $c > 0$ such that for all testing bounds k, the test-outcome sequences S optimal for φ, D, and $\vec{\alpha}$ of length k have the following two properties:*

- *S has at least ck tests of some variable v', but*
- *S has at most $f(k)$ tests of some variable $v \geq_\varphi v'$.*

Then φ exhibits RI.

Proof. Let $P(\varphi, D, \vec{\alpha}, f, c, k)$ denote the statement that for all test-outcomes sequences S that are optimal for φ, D, and $\vec{\alpha}$, there exist variables $v \geq_\varphi v'$ such that S contains $\geq ck$ tests of v' and $\leq f(k)$ tests of v. We now prove that for all φ, D, $\vec{\alpha}$, f, c, and k, $P(\varphi, D, \vec{\alpha}, f, c, k)$ implies the existence of b and g such that φ exhibits RI in the game $G(\varphi, D, k, m, b, g)$. It is easy to see that this suffices to prove the proposition.

Fix φ, D, $\vec{\alpha}$, f, c, and k, and suppose that $P(\varphi, D, \vec{\alpha}, f, c, k)$ holds. Let

$$q^* = \max_{\{S:|S|=k\}} \left|\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}\right|.$$

Since there are only finitely many outcome sequences of length k, there must be some $\epsilon > 0$ sufficiently small such that for all S with $|S| = k$, $|\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}| > q^* - \epsilon$ if and only if $|\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}| = q^*$. Choose the payoffs b and g such that the threshold $q(b, g) = q^* - \epsilon$. We show that φ exhibits RI in the game $G(\varphi, D, k, m, b, g)$.

Let $\mathcal{S}_k = \{S : |S| = k \text{ and } |\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}| = q^*\}$ be the set of test-outcome sequences of length k optimal for φ, D, and $\vec{\alpha}$. If σ is an optimal strategy for the game $G(\varphi, D, k, \vec{\alpha}, g, b)$, the only sequences of test outcomes after which σ makes a guess are the ones in $\mathcal{S}_k$. For if a guess is made after seeing some test-outcome sequence $S^* \notin \mathcal{S}_k$, by Lemma 3.3 and the choice of b and g, that expected payoff of doing so must be negative, so the strategy σ' that is identical to σ except that it makes no guess if S^* is observed is strictly better than σ, contradicting the optimality of σ. So whenever a guess is made, it must be after a sequence $S \in \mathcal{S}_k$ was observed. Since sequences in $\mathcal{S}_k$ are optimal for φ, D, and $\vec{\alpha}$, and $P(\varphi, D, \vec{\alpha}, f, c, k)$ holds by assumption, this sequence S must contain $\geq ck$ test of v' and $\leq f(k)$ test of v.

All that remains to show that φ exhibits RI in the game $G(\varphi, D, k, \vec{\alpha}, g, b)$ is to show that all optimal strategies have positive expected payoff. To do this, it suffices to show that there is a strategy that has positive expected payoff. Let S be an arbitrary test-outcome sequence in $\mathcal{S}_k$. Without loss of generality, we can assume that $\Pr_{D,\vec{\alpha}}(\varphi \mid S) > 1/2$. Let σ_S be the strategy that tests every variable the number of times that it occurs in S in the order that the variables occur in S, and guesses that the formula is true if and only if S was in fact the outcome sequence observed (and makes no guess otherwise). Since S will be observed with positive probability, it follows from Lemma 3.3 that σ_S has positive expected payoff. This completes the proof. □

It is easy to show that all that affects $\Pr_{D,\vec{\alpha}}(\varphi \mid S)$ is the number of number of times that each variable is tested and the outcome of the test, not the order in which the tests were made. It turns out that to determine whether a formula φ exhibits RI, we need to consider, for each truth assignment A that satisfies φ and test-outcome sequence S, the *A-trace* of S; this is a tuple that describes, for each variable v_i, the fraction of times v_i is tested and the outcome agrees with $A(v_i)$ compared to the fraction of times that the outcome disagrees with $A(v_i)$.

In the full paper, we show that whether a formula exhibits RI can be determined by considering properties of the A-

traces of outcome sequences. Specifically, we show that the set of A-traces of optimal outcome sequences tends to a convex polytope as the length of S increases. This polytope has a characterisation as the solution set of an $O(n2^n)$-sized linear program (LP), so we can find points in the polytope in time polynomial in 2^n. Moreover, conditions such as a variable v is ignored while a variable v' that is no more relevant than v is not ignored correspond to further conditions on the LP, and thus can also be checked in time polynomial in 2^n. It follows that we can get a sufficient condition for a formula to exhibit RI or not exhibit RI by evaluating a number of LPs of this type.

Using these insights, we were able to exhaustively test all formulae that involve at most 4 variables to see whether, as the number of tests in the game increases, optimal strategies were testing a more relevant variable a negligible number of times relative to a less variable. Since the criterion that we use is only a sufficient condition, not a necessary one, we can give only a lower bound on the true number of formulae that exhibit RI. In the full paper, we discuss an additional conjecture, the *noise transfer conjecture* (NTC); if it holds, we can establish RI for significantly more formulae.

In the following table, we summarise our results. The first column lists the number of formulae that we are certain exhibit RI; the second column lists the number of additional formulae that exhibit RI if the NTC holds; the third column lists the remaining formulae, whose status is unknown. (Since RI is a semantic condition, when we say "formula", we really mean "equivalence class of logically equivalent formulae". There are 2^{2^n} equivalence classes of formulae with n variables, so the sum of the three columns in the row labeled n is 2^{2^n}.) As the results show, at least 15% of formulae exhibit RI, and this number increases to roughly 30% with the NTC.

n	exhibit RI	NTC $\Rightarrow$ RI	unknown
1	0	0	4
2	8	0	8
3	40	56	160
4	9952	8248	47334

Given the numbers involved, we could not exhaustively check what happens for $n \geq 5$. However, we did randomly sample 4000 formulae that involved n variables for $n = 5, \ldots, 9$. This is good enough for statistical reliability: we can model the process as a simple random sample of a binomially distributed parameter (the presence of RI), and in the worst case (if its probability in the population of formulae is exactly $\frac{1}{2}$), the 95% confidence interval still has width $\leq z\sqrt{\frac{1}{4000}\frac{1}{2}\left(1-\frac{1}{2}\right)} \approx 0.015$, which is well below the fractions of formulae exhibiting RI that we observe (all above 0.048). As the following table shows, RI continued to be quite common. Indeed, even for formulae with 9 variables, about 5% of the formulae we sampled exhibited RI.

n	exhibit RI	NTC $\Rightarrow$ RI	unknown
5	585	313	3102
6	506	138	3356
7	293	63	3644
8	234	30	3736
9	194	10	3796

The numbers suggest that the fraction of formulae exhibiting RI decreases as the number of variables increases. However, since the formulae that characterise situations of interest to people are likely to involve relatively few variables (or have a structure like disjunction or conjunction that we know exhibits RI), this suggests that RI is a widespread phenomenon. Indeed, if we weaken the notion of RI slightly (in what we believe is quite a natural way!), then RI is even more widespread. As noted in Example 4.5, formulae of the form $v_1 \vee (\neg v_1 \wedge v_2 \wedge \ldots \wedge v_n)$ do not exhibit RI in the sense of our definition. However, for these formulae, if we choose the payoffs b and g appropriately, an optimal strategy may start by testing v_1, but if sufficiently many test outcomes are $v_1 \approx F$, it will then try to establish that the formula is false by focussing on one variable of the conjunction $(v_2 \wedge \ldots \wedge v_n)$, and ignoring the rest. Thus, for all optimal strategies, we would have RI, not for all test-outcome sequences (i.e., not in all runs of the game), but on a set of test-outcome sequences that occur with positive probability.

We found it hard to find formulae that do not exhibit RI in this weaker sense. In fact, we conjecture that the only family of formulae that do not exhibit RI in this weaker sense are equivalent to XORs in zero or more variables $(v_1 \oplus \ldots \oplus v_n)$ and their negations (Note that this family of formulae includes v_i and $\neg v_i$.) If this conjecture is true, we would expect to quite often see rational agents (and decision-making computer programs) ignoring relevant variables in practice.

5 Testing as a measure of complexity

The notion of associating some "intrinsic difficulty" with concepts (typically characterised using Boolean formulae) has been a topic of continued interest in the cognitive science community (Vigo 2011; Feldman 2006; Love, Medin, and Gureckis 2004; Shepard, Hovland, and Jenkins 1961). We can use our formalism to define a notion of difficulty for concepts. Our notion of difficulty is based on the number of tests that are needed to guarantee a positive expected payoff for the game $G(\varphi, D, k, \vec{\alpha}, g, b)$. This will, in general, depend on D, $\vec{\alpha}$, g, and b. Actually, by Lemma 3.3, what matters is not g and b, but $q(b, g)$ (the threshold determined by g and b). Thus, our complexity measure takes D, $\vec{\alpha}$, and q as parameters.

Definition 5.1. Given a formula φ, accuracy vector $\vec{\alpha}$, distribution D, and threshold $0 < q \leq \frac{1}{2}$, the *$(D, q, \vec{\alpha})$-test complexity* $\mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}(\varphi)$ of φ is the least k such that there exists a strategy with positive payoff for $G(\varphi, D, k, \vec{\alpha}, g, b)$, where g and b are chosen such that $q(b, g) = q$. □

To get a sense of how this definition works, consider what happens if we consider all formulae that use two variables, v_1 and v_2, with the same settings as in Example 2.1: $\vec{\alpha} = (1/4, 1/4)$, D is the uniform distribution on assignments, $g = 1$, and $b = -16$:

1. If φ is simply T or F, any strategy that guesses the appropriate truth value, regardless of test outcomes, is optimal and gets a positive expected payoff, even when $k = 0$.
2. If φ is a single-variable formula of the form v_1 or $\neg v_1$, then the greatest certainty $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2|$ that is attainable with any sequence of two tests is $2/5$, when $S =$

$(v_1 \approx T, v_1 \approx T)$ or the same with F. This is smaller than $q(b, g)$, and so it is always optimal to make no guess; that is, all strategies for the game with $k = 2$ have expected payoff at most 0. If $k = 3$ and $S = (v_1 \approx T, v_1 \approx T, v_1 \approx T)$, then $(\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2) = 13/28 > q(b, g)$. Thus, if $k = 3$, the strategy that test v_1 three times and guesses the appropriate truth value iff all three tests agree has positive expected payoff.

3. If φ is $v_1 \oplus v_2$, then the shortest outcome sequences S for which $\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2$ is greater than $q(b, g)$ have length 7, and involve both variables being tested. Hence, the smallest value of k for which strategies with payoff above 0 exist is 7.

It is not hard to see that T and F have complexity 0, while disjunctions, conjunctions, and, more generally, majority ("m out of n variables are true") have low complexity. We also completely characterise the most difficult concepts, according to our complexity measure, at least in the case of a uniform distribution D_u on truth assignments (which is the one most commonly considered in practice).

Theorem 5.2. *Among all Boolean formulae in n variables, for all $0 < q \leq \frac{1}{2}$ and accuracy vectors $\vec{\alpha}$, the $(D_u, q, \vec{\alpha})$-test complexity is maximised by formulae equivalent to the n-variable XOR $v_1 \oplus \ldots \oplus v_n$ or its negation.*

Proof sketch. Call a formula φ *antisymmetric* in variable v if $\varphi(A) = \neg\varphi(A')$ for all pairs of assignments A, A' that only differ in the truth value of v. It is easy to check that if a formula is antisymmetric in all variables, it is equivalent to an XOR or a negation of one. Given a formula φ, the *antisymmetrisation φ_v of φ along v* is the unique formula such that $\varphi_v(A) = \varphi(A)$ if $A(v) = T$ and $\varphi_v(A) = \neg\varphi(A[v \mapsto T])$ otherwise. We can show that the $(D_u, q, \vec{\alpha})$-test complexity of φ_v is at least as high as that of φ, and that if $v' \neq v$, then φ_v is antisymmetric in v' iff φ is antisymmetric in v'. So, starting with an arbitrary formula φ, we antisymmetrise every variable in turn. We then end up with an XOR or the negation of one. Moreover, each antisymmetrisation step in the process gives a formula whose test complexity is at least as high as that of the formula in the previous step. The desired result follows. A detailed proof can be found in the appendix of the full paper. □

It is of interest to compare our notion of "intrinsic difficulty" with those considered in the cognitive science literature. That literature can broadly be divided up into purely experimental approaches, typically focused on comparing the performance of human subjects in dealing with different categories, and more theoretical ones that posit some structural hypothesis regarding which categories are easy or difficult.

The work of Shepard, Hovland, and Jenkins (1961) is a good example of the former type; they compare concepts that can be defined using three variables in terms of how many examples (pairs of assignments and corresponding truth values of the formula) it takes human subjects to understand and remember a formula φ, as defined by a subject's ability to predict the truth value of φ correctly for a given truth assignment. We can think of this work as measuring how hard it is to work with a formula; our formalism is measuring how hard it is to learn the truth value of a formula. The difficulty ranking found experimentally by Shepard et al. mostly agrees with our ranking, except that they find two- and three-variable XORs to be easier that some other formulae, whereas we have shown that these are the hardest formulae. Perhaps this is suggesting that there are differences between how hard it is to work with a concept and how hard it is to learn it.

Feldman (2006) provides a good example of the latter approach. He proposes the notion of the *power spectrum* of a formula φ. Roughly speaking, this counts the number of antecedents in the conjuncts of a formula when it is written as a conjunction of implications where the antecedent is a conjunction of literals and the conclusion is a single literal. For example, the formula $\varphi = (v_1 \wedge (v_2 \vee v_3)) \vee (\neg v_1 \wedge (\neg v_2 \wedge \neg v_3))$ can be written as the conjunction of three such implications: $(v_2 \rightarrow v_1) \wedge (v_3 \rightarrow v_1) \wedge (\neg v_2 \wedge v_1 \rightarrow v_3)$. Since there are no conjuncts with 0 antecedents, 2 conjuncts with 1 antecedent, and 1 conjunct with 2 antecedents, the power spectrum of φ is $(0, 1, 2)$. Having more antecedents in an implication is viewed as making concepts more complicated, so a formula with a power spectrum of $(0, 1, 1)$ is considered more complicated than one with a power spectrum of $(0, 3, 0)$, and less complicated than one with a power spectrum of $(0, 0, 3)$.

A formula with a power spectrum of the form $(i, j, 0, \ldots, 0)$ (i.e., a formula that can be written as the conjunction of literals and formulae of the form $x \rightarrow y$, where x and y are literals) is called a *linear category*. Experimental evidence suggests that human subjects generally find linear categories easier to learn than nonlinear ones (Feldman 2006; Love, Medin, and Gureckis 2004). (This may be related to the fact that such formulae are linearly separable, and hence learnable by support vector machines (Vapnik and Lerner 1963).) Although our complexity measure does not completely agree with the notion of a power spectrum, both notions classify XORs and their negations as the most complex; these formulae can be shown to have a power spectrum of the form $(0, \ldots, 0, 2^{n-1})$.

Another notion of formula complexity is the notion of *subjective structural complexity* introduced by Vigo (2011), where the subjective structural complexity of a formula φ is $|Sat(\varphi)|e^{-\|\vec{f}\|_2}$, where $Sat(\varphi)$ is the set of truth assignments that satisfy φ, $f = (f_1, \ldots, f_n)$, f_i is the fraction of truth assignments that satisfy φ such that changing the truth value of v_i results in a truth assignment that does not satisfy φ, and $\|\vec{f}\|_2 = \sqrt{(f_1)^2 + \cdots + (f_n)^2}$ represents the ℓ^2 norm. Unlike ours, with this notion of complexity, φ and $\neg\varphi$ may have different complexity (because of the $|Sat(\varphi)|$ factor). However, as with our notion, XORs and their negation have maximal complexity.

6 Conclusion

We have presented the information-acquisition game, a game-theoretic model of gathering information to inform a decision whose outcome depends on the truth of a Boolean formula. We argued that it is hard to find optimal strategies for this model by brute force, and presented the random-test heuristic, a simple strategy that only has weak guarantees but is

computationally tractable. It is an open question whether better guarantees can be proven for the random-test heuristic, and whether better approaches to testing that are still more computationally efficient than brute force exist.

We used our techniques to show that RI is a widespread phenomenon (at least, for formulae that use at most 9 variables, which certainly covers most naturally-arising concepts for humans). We hope in future work to get a natural structural criterion for when formulae exhibit RI that can be applied to arbitrary formulae.

Finally, we discussed how the existence of good strategies in our game can be used as a measure of the complexity of a Boolean formula. It would be useful to get a better understanding of whether test complexity captures natural structural properties of concepts.

Although we have viewed the information-acquisition game as a single-agent game, there are natural extensions of it to multi-agent games, where agents are collaborating to learn about a formula. We could then examine different degrees of coordination for these agents. For example, they could share information at all times, or share information only at the end (before making a guess). The goal would be to understand whether there is some structure in formulae that makes them particularly amenable to division of labour, and to what extent it can be related to phenomena such as rational inattention (which may require the agents to coordinate on deciding which variable to ignore).

Acknowledgements

We thank David Goldberg, David Halpern, Bobby Kleinberg, Dana Ron, Sarah Tan, and Yuwen Wang as well as the anonymous reviewers for helpful feedback, discussions and advice. This work was supported in part by NSF grants IIS-1703846 and IIS-1718108, AFOSR grant FA9550-12-1-0040, ARO grant W911NF-17-1-0592, and a grant from the Open Philanthropy project.

References

Druzdzel, M. J., and Suermondt, H. J. 1994. Relevance in probabilistic models: "Backyards" in a "small world". In *Working notes of the AAAI–1994 Fall Symposium Series: Relevance*, 60–63.

Feldman, J. 2006. An algebra of human concept learning. *Journal of Mathematical Psychology* 50(4):339 – 368.

Lang, J.; Liberatore, P.; and Marquis, P. 2003. Propositional independence – formula-variable independence and forgetting. *Journal of Artificial Intelligence Research* 18:391–443.

Love, B. C.; Medin, D. L.; and Gureckis, T. M. 2004. Sustain: A network model of category learning. *Psychological Review* 111(2):309–332.

Shepard, R. N.; Hovland, C. I.; and Jenkins, H. M. 1961. Learning and memorization of classifications. *Psychological Monographs: General and Applied* 75(3):1–42.

Sims, C. A. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50(3):665–690.

Umans, C. 1999. On the complexity and inapproximability of shortest implicant problems. In *Proc. of Automata, Languages and Programming: 26th International Colloquium (ICALP '99)*, 687–696. Berlin, Heidelberg: Springer.

Vapnik, V. N., and Lerner, A. Y. 1963. Recognition of patterns using generalized portraits. *Avtomat. i Telemekh.* 24:774–780.

Vigo, R. 2011. Representational information: a new general notion and measure of information. *Information Sciences* 181:4847–4859.

Wiederholt, M. 2010. Rational inattention. In Blume, L., and Durlauf, S., eds., *The New Palgrave Dictionary of Economics (online edition)*. New York: Palgrave Macmillan.