Against the Possibility of a Formal Account of Rationality

Shivaram Lingamneni

January 23, 2013

Abstract

I analyze a recent exchange between Adam Elga and Julian Jonker concerning unsharp (or imprecise) credences and decision-making over them. Elga holds that unsharp credences are necessarily irrational; I agree with Jonker's reply that they can be rational as long as the agent switches to a nonlinear valuation. Through the lens of computational complexity theory, I then argue that even though nonlinear valuations can be rational, they come in general at the price of computational intractability, and that this problematizes their use in defining rationality. I conclude that the meaning of "rationality" may be philosophically vague.

1 Introduction

One task of decision theory (inasmuch as philosophers are interested in it) seems to be providing a formal account of rationality. Such an account should tell the hypothetical "rational agent" what to do; in contraposition, it should constrain his behavior such that any violation of the constraint may be considered "irrational". But it seems that such an account has to avoid two traps. For one, it must separate "rationality" (achievable, ideally susceptible of rule-based description) from "insight" (difficult, presumably not so susceptible). It seems reasonable to expect that rationality might require maximizing expected profit in a game of dice, but not that it might require independently reinventing the theory of special relativity. Rationality seems to be defined as only a limited fragment of our cognitive potential; it does not coincide with our most general notions of "human reason."

Another potential trap is a "no-theory" account of rationality: "just do the right thing." Certainly, yielding the best available action (at least within the context of some constrained space of problems) seems to be a necessary condition for any theory of rationality. However, simply stipulating that the

agent take the best action seems at best a deeply unsatisfying account. For one, it fails to describe how the agent should go about achieving rationality. But furthermore, even in finite or discrete situations where brute force is applicable, it seems to leave the actual content of rationality unexplained.

Call these Pitfall 1 and Pitfall 2. Following a suggestion by Scott Aaronson [2011] that computational complexity theory has significant things to tell us about philosophy — in particular, about issues of human cognition and logical omniscience — I wish to reread a recent exchange between Adam Elga and Julian Jonker in complexity-theoretic terms.[1]

2 Background

I will give an approximate reconstruction of the Elga-Jonker exchange; I must warn the reader that the reconstruction is neither complete nor fully accurate. In the case of Elga, I omit much of the argument, since my concerns differ somewhat from his. In the case of Jonker, I should emphasize that I am reconstructing only an intermediate version of his view — one of his stepping stones, as it were — that I agree with much more than I agree with his final conclusion.

Elga [2010] argues as follows:

1. Consider the following situation (hereafter "Jellyfish Bag"). Imagine that an insane man in the street pulls the following objects out of a bag: a red toothbrush, a live jellyfish, and a green toothbrush. What should your degree of belief be that the next object he will remove will be a toothbrush? There seems to be little or no relevant evidence with which to fix a credence.

2. For this and for a wide class of propositions, it is (apparently) impossible to assign a precise degree of belief ("sharp credence").

3. Consider now the following situation (hereafter "Good Book"). Given an unknown event X, you are offered two bets, A and B. Bet A costs \$10 and wins you \$25 if X is true. Bet B costs \$10 and wins you \$25 if X is false.

4. A rational agent must buy at least one of the bets. A sharp agent maximizing expected utility will buy A if he has $P(X) \geq .4$, and B if he has $P(X) \leq .6$ (both if $P(X) \in [.4, .6]$). But any agent who buys both bets will make a sure profit of \$5, whether X comes true or not. (Elga calls situations of this type "good books", but I will reverse the

[1] I am also following Morton [2004], who applies the P/NP distinction to epistemology and the notion of epistemic virtue.

direction and say that there is a Dutch Book against the agent offering the bets — the offerer is selling bets that will lose him money under every possible outcome.)

5. If an agent's credence in X is sufficiently uncertain — in this case, if it may be in $[0, .4)$ and may also be in $(.6, 1]$ — the agent does not appear to be constrained to buy at least one of the bets. (Elga evaluates and rejects a number of candidate principles that could so constrain the agent.)

6. Elga concludes that in fact, it is irrational to have unsharp credences, since the unsharp agent is not constrained by her credences to take a rational action (recognizing and exploiting the Dutch Book).

Elga does not resolve the apparent contradiction between the Jellyfish Bag and the Good Book; his conclusion appears to be that the rational agent must have a sharp, precise credence for the event of a toothbrush. Jonker [2012] replies (keeping in mind my previous caveats):

1. Implicit throughout Elga's discussion is the assumption that bets be evaluated in a "value-additive" or "linear" way; if you value bet A at $v(A)$ and bet B at $v(B)$, linearity dictates that you value A and B together at $v(A) + v(B)$.[2]

2. In practice, many situations seem to call for nonlinear valuations. In particular, a risk-averse agent can be sublinear. Faced with a small bet, he may value it positively and accept it. But faced with a package containing a million copies of the same bet, he may value it negatively and reject it, because he is afraid to go bankrupt.

3. Similarly, the utility of physical objects can be superlinear. For example, a car with an empty gas tank and a jerrycan of gas are worth more together than apart.

4. A natural perspective on the unsharp agent in the "Good Book" scenario is that her valuations may be superlinear. Due to her uncertain credence in the event X, she may not consider either bet A or bet B worthwhile in isolation. But in that case, she should recognize that together they form a Dutch Book and buy them both.

[2]Elga explicitly restricts himself to agents whose utility is linear in money, but states in a footnote that this is only a notational convenience and his arguments generalize to any "nontrivial utility scale". In light of Ahmed [2006], I think one interpretation of Jonker's counterargument may be that this claim is false, and the Good Book scenario does not generalize to the risk-averse agent, whose utility in money is concave.

5. Model unsharp credences by credence intervals, e.g., [.3, .7]. The following rule accommodates both the sharp agent, who wishes to maximize expected value, and the unsharp agent with interval-valued preferences, who wishes to buy the Dutch Book: choose the action which maximizes (over all possible actions) the minimum expected utility (over all credence values in the interval). Call this rule "maximin expected utility", or MMEU.

Jonker's ultimate rejection of MMEU is motivated by the simplistic nature of the interval-valued credence model, also by the way it seems to conflate unsharpness with risk aversion. Certainly, MMEU (as a maximin principle) does not seem to describe all possible rational responses to unsharpness. An optimistic agent with "nothing to lose" (perhaps identifiable with the risk-seeking agent) might prefer a maximax principle, choosing the bets with the greatest possibility of gain, no matter how unlikely that possibility is. In between these two, we have the possibility of collapsing the unsharp agent into a sharp, risk-neutral agent, whose first-order credences are precisely the expected values of her second-order credence distribution. I do not wish to argue that the maximin response to unsharpness is the only response, or a universally applicable one. My discussion will depend only on these contentions:

1. MMEU is a successful counterexample to Elga's claim that no decision principle constrains the unsharp agent to buy at least one bet from the Good Book. The maximization in MMEU is over four possibilities: buy neither, buy A, buy B, or buy both. The minimum expected utility from buying neither is \$0, and the minimum expected utility from buying both is \$5, so it is impossible for the agent to buy neither.

2. In at least *some* situations, MMEU is in fact the ideally rational response to unsharp beliefs. This depends on two subclaims:

   (a) MMEU successfully models the risk-averse response to unsharp credences, and in the extremal case, an attitude I will call "absolute risk aversion": the unwillingness to countenance any possible loss.

   (b) There exist situations in which risk aversion (even absolute risk aversion) is the ideally rational response.

Claim (1) is, I think, evident. For Claim (2a), see proposition 2, but the intuitive justification is simply that the MMEU-agent always considers the worst possible situation consistent with his beliefs.

Claim (2b) is potentially more controversial, but the rationality of risk aversion has been extensively discussed by philosophers and economists. I

will offer just one scenario in which absolute risk aversion seems justified: bankruptcy. Imagine that an agent is offered a package of bets, the expected value of which is high, but which admits at least one possible outcome in which she loses all her money. Furthermore, she is confident that in the future, she will be offered a stream of Good Books — as long as she has the money to buy them in the first place. Alternately, imagine that in the agent's society, the penalty for bankruptcy is to be sold into a lifetime of indentured servitude.

This suggests a computational rereading of Elga's argument. Elga has shown that there exist Dutch Books such that for an unsharp agent, every individual bet of them is not worthwhile in isolation. Thus, rationality requires the unsharp agent to have the ability to recognize Dutch Books; otherwise she risks missing out on sure gain. However, linear (or value-additive) decision principles such as expected utility maximization (hereafter EUM) are not powerful enough to recognize Dutch Books. The unsharp agent must replace EUM with something like MMEU.

I am a frequentist, so I am fully convinced of the necessity of unsharp credences. I consider them the natural and correct response to many cases when the reference class for an event (such as "toothbrush") is ambiguous or inadequate. (Hájek [2007] argues convincingly that Bayesianism does not provide easy answers in these cases either.) So it might seem that I am committed to MMEU or something like it as an account of decision-making over unsharp credences. But going beyond frequentism per se, I am generally skeptical of quantitative accounts of belief and rationality, orthodox Bayesian or otherwise — so I will try and dispute MMEU as well.

The most basic objections to the MMEU picture that occur to me are methodological. Why should the unsharp agent be able to quantify the exact nature of her unsharpness? Saying that imprecise beliefs are described by precise intervals of credence (or precise second-order belief distributions) seems to be introducing a false precision — and once we have abandoned the Ramseyan argument that precise credences can be elicited by measuring the agent's propensity to bet, it is unclear how we can measure these higher-order beliefs, no matter how we represent them. Perhaps an unsharp belief interval can be interpreted as a bid-ask spread on the bet in question, but a general concern about false precision still remains. These objections lead rapidly into abstract concerns and touch on a longstanding controversy in the philosophy of probability, and I will not discuss them further here.

A second basic objection, and one mentioned by Jonker, is that MMEU does not seem like the last word in decision principles. Expected utility maximization neatly avoided Pitfall 1; it gave a simple condition for action against betting books, namely evaluating each bet in isolation. Now that we have accepted this more elaborate principle, is the door open for us to

require more and more complex criteria for rationality? This is the objection I intend to pursue formally here. In order to do it, I will introduce some notions from computational complexity theory.

3 Computational complexity and philosophy

Unlike recursion (or "computability") theory, in which the main objects of study are problems that cannot be solved by any computer, computational complexity theory studies the relative hardnesses of problems that computers can solve. Speaking very loosely, the problems we are ordinarily accustomed to solving with computers (arithmetical operations, sorting, shortest paths in maps, etc.), are in the complexity class P, meaning that they can be solved within a time that is polynomial in the size of the input.

There is a natural class of prima facie harder problems, known as NP. Intuitively, problems in NP have the following form: they can be computed by an algorithm that "guesses" a solution from an exponential search space, then verifies it in polynomial time. The canonical problem of this type is SAT, or Boolean satisfiability: the question of whether a formula of propositional logic is true under some assignment of truth values to the atoms. Checking whether a particular assignment satisfies the formula is easy (i.e., polynomial-time), but given n atoms, there are 2^n possible assignments overall — thus, the brute-force solution to SAT requires time at least exponential in the size of the input. P is clearly contained in NP. Although it is strongly suspected that in fact P $\neq$ NP, this has not been proven; it is considered one of the major unsolved problems in contemporary mathematics.

The "hardest" problems in NP are called *NP-complete*. Their defining characteristic is that every problem in NP is reducible to them, so if any of them were discovered to be in P, it would imply P = NP. (Specifically, for any NP-complete problem Q, there is a polynomial-time many-one reduction, or Karp reduction, from any problem in NP to Q.) Problems outside NP may be *NP-hard*, intuitively, at least as hard as NP-complete problems. (Formally, Q is NP-hard if there is a polynomial-time Turing reduction, or Cook reduction, from any problem in NP to Q.)

SAT is NP-complete. It has subproblems called k-SAT that are also NP-complete:

Theorem 1. *A* literal *is a propositional formula of the form a or $\neg a$, i.e., a positive or negated atom. Let a k-*ary disjunction *be a disjunction of k literals; likewise for k-*ary conjunctions*. For $k \geq 3$, the problem k-SAT of determining the satisfiability of conjunctions of k-ary disjunctions is NP-complete.*

3.1 The exponential time hypothesis

The *exponential time hypothesis* (hereafter "ETH") of Impagliazzo and Paturi [2001] is slightly stronger than P $\neq$ NP. It has various forms, but in general it says that the hardest NP-complete problems cannot be solved in subexponential time, i.e., $2^{o(n)}$. For example, $O(2^{\sqrt{n}})$ is considered subexponential under this definition, but $O((\sqrt{2})^n) = O(2^{0.5n}) \in 2^{O(n)}$ is not, even though both are asymptotically faster than $O(2^n)$. As with P $\neq$ NP, the ETH is unproven but widely believed.

Conjecture 1 (Exponential time hypothesis)**.** *For each k, let s_k be the infimum (greatest lower bound) of the set of reals $\{\delta \mid$ k-SAT is solvable in $O(2^{\delta n})\}$. For $k \geq 3$, $s_k > 0$.*

We have known upper bounds on s_3, the best due to Moser and Scheder [2010]:

Theorem 2. *3-SAT is solvable in $O((\frac{4}{3} + \epsilon)^n) \approx O(2^{0.416n})$, for arbitrarily small $\epsilon > 0$. Consequently, $s_3 \leq 0.416$.*

So there are solutions to 3-SAT that asymptotically outperform brute force, despite still being exponential.[3] But in the general case, we have a (slightly stronger again) conjecture by the same authors:

Conjecture 2 (Strong ETH)**.** $\lim_{k\to\infty} s_k = 1$.

The Strong ETH says that for larger and larger values of k, the optimal solution of k-SAT regresses progressively to the brute-force $O(2^n)$ solution that tests all possible assignments.

I will idiosyncratically refer to the problem of deciding whether a propositional formula is a tautology as VAL (for "validity"). The specific form of VAL where the formulae are 3-ary conjunctions of positive or negated atoms (by analogy with 3SAT) will be called 3VAL. VAL and 3VAL are unlikely to be in NP (they naturally fall in co-NP instead), but since they are the complement problems of SAT and 3SAT, they are as hard:

Proposition 1. *VAL and 3VAL are NP-hard, and exponential lower bounds on SAT and 3SAT (respectively) apply to them as well, i.e., under the exponential time hypothesis, they require exponential time.*

Proof. Assume a subexponential algorithm for 3VAL. Take an instance of 3SAT of the form:

$$(a \vee b \vee \neg c) \wedge (\neg b \vee d \vee e) \ldots$$

[3] In passing, although the ETH only talks about deterministic algorithms, the best known randomized algorithms for k-SAT are also exponential.

and compute its negation:

$$(\neg a \wedge \neg b \wedge c) \vee (b \wedge \neg d \wedge \neg e) \ldots$$

Apply the algorithm for 3VAL, then invert the answer (a formula is satisfiable iff its negation is not a validity). The transformation is polynomial-time, so this is a Cook reduction from 3SAT to 3VAL. Moreover, the transformed formula has the same number of variables and clauses as the original, and we invoked the oracle exactly once, so we have a subexponential algorithm for 3SAT, which contradicts the ETH.

The proof for SAT and VAL is similar. □

3.2 The tractability criterion

Complexity theory gives a natural (partial) formalization of Pitfall 1. Rational agents have limits on their computational power; they cannot be expected to perform arbitrarily difficult optimizations. So it seems natural to propose the following *tractability criterion*:

Criterion 1. *Assume $P \neq NP$. Any decision principle proposed as a constraint on rational agents must have an algorithm in P.*

I hold this condition to be necessary, but not sufficient — but I will postpone discussion of the converse principle until section 5.

How much you believe the tractability criterion will depend on how much you believe logical omniscience is a problem for Bayesian rationality. Personally, I believe that it is a very serious problem, and that it has not been adequately addressed. In particular, I am aware of two approaches to the problem, neither of which seems relevant to the present problem.

1. Formal systems such as Garber [1983], which attempt to relativize Bayesian rationality to an agent who is ignorant of some logical truths. Whether these systems succeed is unclear to me. But whether they do or not, they are not addressing Elga's question of what the rational agent is in fact obligated to know. In the extremal case, they would allow the agent not to know that $X \vee \neg X$ is a tautology, i.e., that the Good Book pays off under every outcome.

2. The approach of Levi [1997], who holds that all agents are obligated by the constraints of ideal, logically omniscient rationality, even if the obligations are not realistically achievable. Levi argues that the rational agent is obliged to bring himself as close as possible to ideal rationality, by seeking out "training and education", "appropriate therapies", and "prosthetic devices."

> I think that this position is simply not persuasive in the face of computational complexity theory. If the ETH is true, there are problems that the rational agent can formulate in minutes, but which he cannot be assured of solving before the heat death of the universe — no matter what prostheses he invents. The naive ideal of rationality is not merely unachievable, it is also unapproachable. Levi's position is convincing if we believe the computational demands of rationality increase in a manner commensurate with the problems we wish to solve; then we might realistically hope that our abilities could keep pace with those demands. But the complexity-theoretic consensus is that this is not so.

In passing, Hacking [1967] gives an account of personal probability without logical omniscience that explicitly incorporates the cost of computation and appears to sidestep my first objection. But he glosses over details that I think are important and not readily filled in. I will respond to some of Hacking's arguments in sections 4 and 6.2.

But even if we gloss over the problem of intractability entirely, I have a different objection against unrestricted decision principles — one related to the other pitfall.

3.3 The guidance criterion

I hold that the exponential time hypothesis almost precisely captures the intent behind Pitfall 2. If it is true, then there are search problems for which no help is possible; every successful algorithm is, in the worst case and modulo constant-factor speedups, a brute-force search in disguise. Intuitively, a brute-force search has no philosophical content; it does not tell us anything about the problem it solves. So this motivates my proposed *guidance criterion*:

Criterion 2. *Assume the exponential time hypothesis. A decision principle that requires the agent to solve SAT or 3SAT is* content-deficient*, because it requires a brute-force search in the worst case — it offers the agent no meaningful* guidance*.*

The notion of content I am invoking here needs to be clarified. A decision principle that requires a brute-force search has a certain kind of content — at the very least, it says what to search for! But in the sense that a decision principle should help you decide, it provides the minimum possible help. One way to understand this is that its content is purely definitional: it supplies only a definition of value or a preference ordering, but not instructions on how to achieve or satisfy it. Informally, it tells you what "best" means, then tells you, "just pick the best thing."

Another necessary clarification has to do with heuristics. Aaronson notes that since most complexity theory focuses on worst-case analysis, applications of it to philosophy are subject to a general objection: perhaps the worst case is not philosophically relevant, and what really matters is the average or typical case? In practice, many heuristic algorithms for SAT ("SAT solvers") have sophisticated *algorithmic* content and substantially outperform brute force on typical real-world inputs. I am not saying that these algorithms are trivial, but rather that they cannot rescue a decision principle reliant on them from content-deficiency.

Why should a *worst-case* regression to brute force mean that a decision principle lacks content? I have two arguments, one I think is good and one I think is middling:

1. A key function of rational decision principles is to prevent you from being exploited by malicious agents. (This is, in fact, the standard Dutch Book argument for belief in the probability axioms — if you violate them, someone can construct a Dutch Book against you.) But if you rely on an imperfect heuristic, the malicious agent can theoretically feed you exactly the problem instances that break it.[4] Regressing to brute-force in the face of a malicious adversary is, with respect to this requirement, as bad as regressing all the time.

2. Consider someone who believes a SAT-solving decision principle is content-deficient over the worst-case problem instances, but not in general. This person owes us an explanation of where the principle succeeds and fails. Saying that the principle only has meaning over some unspecified subset of the problem space is an unacceptable retreat from the original notion of decision principle.

Another important qualification is that, as mentioned above, 3SAT can be solved deterministically in $O(c^n)$ for $c < 2$, which is an asymptotic improvement over the brute-force $O(2^n)$. But I do not think this can be construed as offering a sufficient level of guidance. Notice that $O(c^n)$ is $O(2^{\log_2 c \cdot n})$ — so the running time is equivalent to "shrinking" the problem by a constant factor of $\log_2 c$, then performing a brute-force search on the remaining space. Given a constant $d < 1$, $2^{O(dn)}$ does not seem qualitatively less brute-force than $O(2^n)$, for the same reasons that we generally neglect constant factors when doing asymptotic analysis. In any case, under the Strong ETH, SAT and k-SAT for arbitrary k are necessarily brute-force in the original sense.

[4]This phenomenon is observed in practice. An algorithm with a high worst-case running time is a potential security vulnerability for a network service such as a website, because sending it the pathological inputs can result in a denial-of-service attack. For a recent instance "in the wild", see `http://bugs.python.org/issue13703`.

4 Technical results

A disclaimer: every result in this section is more or less trivial. The least trivial result, hardness of DUTCHBOOK, has been proved already in greater generality (see Paris [1994]). I include proofs so that the arguments can be compared at every step with the intuitions motivating the hypotheses, in hope of showing that the results are founded on essential aspects of the problem, rather than accidents of the mathematical formalism.

The natural generalization of Elga's Good Book scenario is the problem of betting on books of propositional formulae. In the case we are interested in, these propositional formulae consist of Boolean combinations of atomic events, for some set of n atoms. Furthermore, we will require that the agent regard all 2^n possible combinations of these atomic events as logically possible — equivalently, that the agent is not aware of any logical implications among the events. If such a thing seems outlandish, consider the atoms $\{f_i \mid 1 \leq i \leq 31\}$, where f_i denotes the event that I will eat falafel for lunch on the ith day of January. The f_i may not be probabilistically independent, but they certainly seem to be logically independent. There really are $2^{31} = 2147483648$ distinct possibilities for my lunches.

The concept of Dutch Book is also in need of some formal clarification. Hacking [1967] points out that in cases where the bookie knows more facts than the bettor, the result may be a trivial Dutch Book. For example, when the bettor pays \$0.5 for a \$1 bet that the quarter will land tails, but unbeknownst to him the bookie has provided a two-headed quarter, the bettor has been Dutch Booked in the sense that he loses money under every possible outcome. I think this is not properly in the spirit of the original definition of Dutch Book. The relevant quantification is over all outcomes the bettor perceives as logically possible — and this includes the excluded possibility of tails. It seems that it should also include events to which the agent assigns probability 0 — for example, if we model the heights of men with a real-valued normal distribution, we consider it logically possible for a man to be exactly six feet tall, even though this occurs with zero probability. For this reason, I will define a Dutch Book as one that pays off over every outcome in the agent's state space, whether or not the agent assigns the outcome nonzero probability.

Definition 1. *A* decision principle *for betting books is an algorithm that takes in a book of propositional bets B and a representation C of the agent's credences, then outputs what bets the agent should buy. (EUM and MMEU are decision principles in this sense.)*

Definition 2. *Let* DUTCHBOOK *be the following decision problem. Given a book of propositional bets over n atoms, does there exist a package of bets that yield a profit under all 2^n outcomes?*

Proposition 2. *DUTCHBOOK is Karp-reducible to MMEU over the class of all propositional books. (In other words, MMEU is harder than DUTCHBOOK, or "contains" it.)*

Proof. Given a book of bets we intend to test for Dutchness, we construct an agent whose degree of belief in every atom is uncertain between 0 and 1, inclusive. We apply MMEU to this agent; the agent will buy a package of bets if and only if they are a Dutch book. (Intuitively, the agent who is completely unsure of every proposition can only justify betting when victory is assured, no matter the outcome.)

If we wish to avoid applying MMEU to agents with extremal beliefs (personal probabilities of 0 and 1), we can simply substitute beliefs that are sufficiently close to 0 and 1. I omit a detailed proof, but it should suffice to replace 0 with an ϵ satisfying

$$0 < \epsilon < \frac{1}{\sum_i w_i} \cdot \frac{1}{\gcd(w_1, w_2, \ldots w_n)}$$

where the w_i are the payoffs of the various bets. Replace 1 with $1 - \epsilon$; both of these are clearly computable in polynomial time. □

Proposition 3. *DUTCHBOOK is Karp-reducible to any principle that computes an optimal decision for a completely risk-averse agent.*

Proof. As above. The risk-averse agent can only buy Dutch books. (As we mentioned previously, the fact that the risk-averse agent assigns probability 0 to an outcome does not exclude it from the definition of Dutch Book.) □

Proposition 4. *DUTCHBOOK over the class of books consisting of bets on atoms or negated atoms is in P.*

Proof. The book consists entirely of prices on bets on A and $\neg A$, for various logically independent atomic propositions A. Sort the bets by which proposition they describe. If the book gives two distinct prices for some A, there is a Dutch Book (sell the bet at the higher price and buy at the lower); if it prices $\neg A$ at something other than 1 - Price(A), there is a Dutch Book (either buy both bets or sell both bets). If neither of these is true for any A, the bookie's beliefs are Kolmogorov consistent and there is no Dutch Book against him. □

Elga's original example falls into this class of books. But if we generalize to arbitrary propositions, the problem becomes harder — the definition of Dutch Book quantifies over all 2^n possible outcomes, and we can harness this quantification to solve hard problems.

Proposition 5. *DUTCHBOOK over the class of books consisting of arbitrary propositions is NP-hard. Furthermore, under the Exponential Time Hypothesis, it requires at least exponential time.*

Proof. Fix a propositional formula φ. Apply DUTCHBOOK to the book consisting of a single bet on φ, priced at \$0.5 and paying \$1. This book is Dutch if and only if φ is a propositional validity; this is a Karp reduction of VAL to DUTCHBOOK. □

This seems somewhat cheap. Naturally, if we confront our agent with an arbitrarily complex propositional formula, we might expect bewilderment. In particular, it is not clear that the expected-utility-maximizer can do better on this problem. But consider what happens when we replace VAL with 3VAL.

Definition 3. *Let the* 3-Books *be the class of propositional books where every proposition is a 3-ary conjunction, i.e., of the form* $(p_1 \wedge \neg p_2 \wedge p_3)$.

Proposition 6. *DUTCHBOOK over the class of 3-Books is NP-hard. Furthermore, under the ETH, it requires at least exponential time.*

Proof. Fix an instance of 3VAL, i.e., a formula φ that is the disjunction of n clauses of the abovementioned form. Construct the following book: for each clause, offer a bet, priced at \$1, that pays $\$(n+1)$ if the clause comes true.

This book is Dutch if and only if φ is a validity. If φ is a validity, then under every possible outcome, φ must be true, so at least one of its disjuncts must be true, so buying every bet costs $\$n$ and pays at least $\$(n+1)$, for a sure gain of at least \$1. Conversely, if φ is not a validity, then under the truth assignment that makes it false, no bet pays off and every package loses money. This is a Karp reduction of 3VAL to DUTCHBOOK. □

Now we can see the computational advantage of sharpness:

Proposition 7. *The sharp, risk-neutral Bayesian agent with an oracle for personal probabilities can evaluate 3-Books in polynomial time. Performing this evaluation requires only polynomially many beliefs.*

Proof. The agent examines every bet in isolation, computes its expected payoff against her personal probability, and accepts the bet iff the payoff is positive. This is linear time, or $O(n)$. The number of 3-ary propositions she can be asked to bet on is bounded above by $\binom{2n}{3}$, which is $O(n^3)$. □

The assumption of an oracle coincides with the intuition that a Bayesian agent has roughly direct access to her level of credence in simple propositions. For a specific example, consider the agent who assumes conditional independence of all the atoms, then applies the principle of indifference to each one. Her credence in each 3-ary proposition is then $(\frac{1}{2})^3 = \frac{1}{8}$. Less trivially, the agent could model the propositions as a Markov chain or a polytree Bayesian network, both of which admit polynomial-time algorithms for computing the probabilities of conjunctions. However, the assumption is not entirely unproblematic. I discuss potential failures later.

Parenthetically, the relationship between risk aversion and NP-hardness is discussed in the economic literature. In particular, Ahmed [2006] proves the NP-hardness of a certain kind of stochastic optimization that accounts for risk, and Bertsimas et al. [2010] prove the NP-hardness of a minimax criterion for risk aversion. I am unclear on exactly how their results relate to mine — but I wouldn't be surprised if everything in this section is an elementary corollary of their work.

5 The dilemma

Assume either that it can be rational to have unsharp credences, or that absolute risk aversion can be rational. Assume P $\neq$ NP and the Exponential Time Hypothesis, and let DEC be a proposed decision principle in the sense of definition 1. I argue that DEC is caught on at least one horn of the following dilemma:

1. DEC runs in polynomial time.
   (a) In general, DEC is not powerful enough to identify Dutch Books.
   (b) Specifically, consider the following class of decision problems: 3-Books with n clauses, viewed by either by a absolutely risk-averse agent or an unsharp agent whose credence interval for each atom is $[p, 1-p]$ for $p < \frac{1}{n+1}$. There must exist instances of this class where DEC does not yield the optimal action (buying the book if and only if the clauses form a validity).
   (c) A generalization of Elga's argument applies to DEC: DEC is an inadequate condition on rationality, because it does not constrain unsharp and risk-averse agents to take a rational action (exploiting a Dutch Book).
2. DEC is powerful enough to identify Dutch Books.
   (a) DEC must (in general) run in exponential time.
   (b) DEC has fallen into Pitfall 1: it is computationally intractable and therefore represents an unachievable standard of rationality.
   (c) DEC has fallen into Pitfall 2: since it instructs the agent to perform an exponential-time brute-force search, it is content-deficient.

EUM is gored by the first horn (it is tractable but inadequate), and MMEU by the second (it is adequate but intractable). And we could construct degenerate principles that have neither desirable property. But by proposition 6, no principle can have both.

Note that in order to generalize ("scale", perhaps) the problem to larger books, we required the existence of agents with increasingly wide unsharp belief intervals: if the agent's interval does not go below $\frac{1}{n+1}$, then she can justify buying the bets in the 3-Book based on their expected utility alone. This is a problem, but not, I think, a serious one. In particular, these thresholds are shrinking only reciprocally, while the time complexity of the 3-Books is increasing exponentially; it's easy to imagine someone whose dubiety extends from 1% to 99%, but 2^{100} is already astronomically large.[5]

This dilemma purports to show the failure of every possible decision principle for betting books. How might one go about denying this conclusion? Here are all the possibilities I can think of:

1. Deny the rationality of sharp credences (Elga's stated position). This involves giving a sharp credence for the Jellyfish Bag, or at any rate describing how to obtain such a credence. It also involves denying the rationality of risk aversion (or at least risk aversion beyond certain thresholds). I do not think this is an attractive option.

2. Deny the validity of Elga's Good Book argument, since it appears to constrain the rational agent to solve NP-hard problems. I think this is unattractive because Elga's argument, in its original form, is very persuasive:

   > I can only ask you to vividly imagine a case in which an agent rejects both bets A and B. Keep in mind that this agent cares only about money (her utility scale is linear), that she is certain in advance what bets will be offered, and that she is informed in advance that her state of opinion on the bet proposition will remain absolutely unchanged throughout the process. I invite you to agree with me that this agent has exhibited a departure from perfect rationality.

   Even if the generalized problem is intractable, that does not seem to undermine Elga's invitation; any notion of rationality that does not constrain the agent to buy A or B is surely too weak. It seems like Elga's Good Book argument is valid for his original problem instance (metaphorically, the validity of $X \vee \neg X$) — it just may not be valid in all the cases I am applying it. So this leads us to a different class of objections, those attacking my extrapolation from Elga's argument.

3. Reject (both) the tractability and guidance criteria. I have argued for their acceptance in sections 3.2 and 3.3.

[5] Actually, given the enhanced running time for 3SAT in Theorem 2, we should multiply by $c < .416$ and use, e.g., the agent with credence interval $[.4\%, 99.6\%]$. But as discussed before, this is only a constant-factor change.

4. Reject the generalization from Elga's 1-Book (bets on X and $\neg X$) to 3-Books, since the problem for 1-Books is in P and the problem for 3-Books is not. This is the natural position for someone who considers P the definition of tractability, i.e., affirms both criterion 1 and its converse.

   As attractive as such a position might seem, I do not think it is justified. The problem is that small 3-Books are just as easy as Elga's small 1-Book — and Elga's argument that the rational agent must solve the 1-Book makes no reference to P and NP, but merely to the intuitive plausibility of recognizing the Good Book. It seems that what the argument really hinges on is the real-world tractability of the problem: the fact that it can be solved without much computational power, i.e., without thinking very hard.

   Does P capture real-world tractability? In fact, the engineering consensus is that it is neither a necessary nor a sufficient condition. An exponential-time algorithm may be fast in practice if the input sizes are small, but a polynomial-time algorithm may be unusable if the exponent (e.g., $O(n^{1000})$) or the constant factor is too high. I think that to date, the best attempt to model real-world tractability has been the complexity-theoretic one, and its main result has been negative: tractability is not characterized by any complexity class.

   It would, in my opinion, be entirely reasonable to say that a rational agent with unsharp credences is obligated to run MMEU on small problems, in the case of an ordinary human being, perhaps up to systems with 16 distinct states. And I think Elga's argument is persuasive in exactly the situations where the problem can be solved "fast-in-practice". This leads us to my preferred resolution of the dilemma:

5. Reject a strict interpretation of my algorithmic formulation — abandon the idea that a decision principle must solve all 3-Books, or that it must always run in polynomial time. I will retreat from the universal applicability of the guidance and tractability criteria, and admit the Good Book argument over a problem space with vague boundaries that does not coincide with any natural mathematical definition. But I will not abandon the attempt to analyze rationality using computational formalism; I think the relevance of computation extends well beyond the betting situations discussed so far.

6 Rationality and computation in general

6.1 Combinatorial optimization as a fact of life

NP-hard problems have a way of intruding into situations where one might not expect them. Robert [2007] shows that they can arise naturally in Bayesian statistics, when estimating hyperparameters. But they abound in real-life situations as well. For example, the well-known Travelling Salesman Problem models the situation faced by a person who wishes to plan a trip that will visit n cities exactly once, finally returning to his starting place. Finding the minimal such tour turns out to be NP-complete.

There is a slippery slope here. Let's say the the agent is an actual traveling salesman, and real money is riding on how fast he can cover his cities. We can adapt our notion of decision-theoretic rationality to the specific, constrained problem he faces: choosing a route. It seems that rationality should constrain him to avoid certain pathological routes (for example, the tour that always visits the most distant unvisited city, which will be suboptimal on nontrivial inputs). But does rationality also constrain him to find the optimal solution? It seems fairly clear that the answer is "no", because picking a suboptimal tour and getting back to work is preferable to waiting indefinitely for the optimal solution. In practice, he can try a fast-in-practice heuristic technique such as branch-and-bound; if it doesn't finish and further delay would lose him business, he can switch to a polynomial-time approximation algorithm like Lin-Kernighan or Christofides. But does rationality entail knowing these algorithms as well? That seems absurd. There is nothing particularly natural, obvious, or universal about them.

It is important to note that although an exact polynomial algorithm for one NP-complete problem would yield exact polynomial algorithms for all of them, the same is not true for approximation algorithms; guarantees such as "yielding the optimum result to within a constant factor" are not in general preserved by polynomial-time reductions. For example, the NP-complete problem of set cover (which plausibly abstracts a situation that arises in personnel hiring) is believed to be inapproximable to within a logarithmic factor. So even within the class of NP-complete problems, it does not suffice to teach the rational agent a single approximation technique — messy domain-specific knowledge seems to be required. If a decision-theoretic notion of rationality applies here, it is a Balkanized one.

Tsang [2008] calls this idea *computational intelligence* — researching new heuristics enhances your ability to be rational. More generally, the economic literature on bounded rationality attempts to relativize the notion of rationality to agents with limited computational power, going back to Herbert Simon's idea of the satisficing agent. However, this literature has not yielded anything like a natural, domain-universal characterization of bounded ratio-

nality. Given the difficulties outlined above, this is unsurprising.

6.2 Hacking and time management

As I have previously alluded to, Hacking [1967] makes a good case for the idea that Bayesian rationality relativizes gracefully to agents who are not computationally omniscient. Ultimately, his argument rests on the idea that agents can measure the costs of thinking and weigh them against the benefits. An agent is then obligated to discover logical truths, or solve optimization problems, in exactly the cases where it is worth her while. This is a persuasive argument, and one I more or less agree with. But I do not think it is precise, and I think there are considerable obstacles to making it precise.

Hacking analyzes the cost of thinking by modeling all reasoning as repeated applications of modus ponens, then charging the agent a fixed price (originally \$0.25) for each application. Now, given a computation C, the agent should determine her expected utility U_0 from acting without knowing the result of C, then her expected utility U_C from acting with that knowledge. She should then do C iff $U_C - U_0 \geq \text{cost}(C)$.

I find this problematic, but not because of the assimilation of all computation to logical inference; we can substitute Turing machine steps, CPU cycles, or seconds of real time as the units of computational effort. The problem is that this does not seem to describe all the situations in which we think. Although there are certainly contexts in which the cost of computation scales linearly with the number of steps (metered pay-by-the-hour cluster computing is one example), a more typical situation seems to be the following: the agent can compute for free, as long as she finishes by a deadline, after which the answer is no longer useful. In Hacking's language, this agent's marginal cost of computation is zero up until the deadline, but afterwards it is high enough to cancel out all benefits — her utility in computational power is nonlinear. A general account of the cost of reasoning must, at the very least, accommodate a variety of these utility curves.

Determining U_C and U_0 seems to require having accurate priors on the space of problems. For example, over the class of 3-Books described in section 5, U_C would depend on the prior probability that the book is Dutch (since it is worth \$1 if it is and nothing if it is not). But I think a more interesting difficulty is this: even if U_C and U_0 are known and $\text{cost}(C)$ is linear in computation time, the agent trying to follow Hacking's prescription still has to know how long C will take. If she has a polynomial-time algorithm for C, then she knows an upper bound. But if C is NP-hard, then she has to account for the probability that she will not be able to solve C in time. Problems of this kind are studied in *average-case complexity* (see Bogdanov and Trevisan [2006] for a survey); the field is in its early stages, but a common theme of its results is that they are sensitive to the choice of distribution over the space

of problems.

To make Hacking's idea precise for NP-hard problems, we might imagine fixing a particular problem, then fixing a particular heuristic algorithm for it that is exponential in the worst case. Then what we would want is a polynomial-time algorithm that takes in instances of the problem and produces estimates of the time needed to solve them, in the form of probability distributions over running times.[6] To the best of my knowledge, no estimation scheme like this is known, but that does not rule out the possibility that one could exist; perhaps in some specific cases, Hacking's platonic ideal of rational time management can be rigorously realized. But as in the previous section, it seems implausible that a single nice characterization spanning all problems will emerge.

6.3 Search problems and creativity

The car-and-gas problem, suggested by Jonker, is an example of real-life nonlinearity. Let's say you want to go to the beach. A car with an empty gas tank and a jerrycan of gas, taken separately, are unhelpful. But combine them (appropriately) and all is well.

I think this scenario is like Elga's Good Book: it only looks tractable because it is so small. Reconstructing the reasoning:

1. A rational agent should realize that neither the empty car nor the gas can get to the beach in isolation.

2. A rational agent should realize that the car and the gas can be combined and driven to the beach.

3. In contraposition: an agent who does not realize this has violated rationality.

What happens if we generalize this problem to n objects? The agent's problem is now to identify useful subsets of the objects, and there are 2^n such subsets. I now have 2^n opportunities to claim that the agent is irrational! But more to the point, it seems that now the agent is no longer simply *scanning* the available items, but *searching* for the right combination of items. The fact that (for example) the magnifying glass can be used to ignite the phone book and send a smoke signal must *occur* to him.

Is the rational agent necessarily someone of limitless ingenuity, seeing all the possible uses for the objects around him? Once we retreated from a linear

[6]For example, Coarfa et al. [2000] note that if the ratio of clauses to variables in a 3-SAT problem is less than 4.26, heuristic algorithms can usually find a satisfying assignment quickly, and if it is greater, they can usually tell quickly that it is unsatisfiable. The genuinely hard problems live, for the most part, in a small interval around 4.26. Since this ratio is very easy to measure, this provides the beginning of such an algorithm.

(or "local") notion of utility, focused on the independent usefulness of each item, and moved to a nonlinear model focused on the interactions between items, we opened a Pandora's Box of possibilities for recognizing utility — much like the wealth of possibilities we confront with real objects in real situations.

I wish to emphasize that this version does not have a natural formalization as an NP-hard problem; specifically, it is difficult to imagine a presentation of the problem that does not give all utilities for subsets of the items up front, at which point picking an optimal member is trivially $O(n)$.[7] But informally, I think this example hints at Avi Widgerson's idea [2009] that the P/NP distinction corresponds to the divide between ordinary thinking and creative insight — that P corresponds naturally to straightforward problems of checking or verification, but the additional hardness of NP corresponds to the difficulty of invention or insight. It is "hard", in some informal sense of the word, to perceive potential relationships among objects, harder than it is simply to recognize them as individuals. As n grows large, picking out combinations of items becomes a creative act, beyond the scope of simple "rationality".

7 Rationality — a vague notion?

Elga's argument hinges on the interpretation of "rationality" as a constraint or obligation — a standard agents must conform to. But "ought" implies "can", and the maximal notion of rationality is degenerate; it entails the obligation to do impossible things without number or limit. Elga's Good Book argument is believable because it does not appeal to such an an expansive definition, only a small fragment. What happens if we try to clarify the notion Elga is invoking? What could rationality really be?

I do not think rationality can consist in any specific principle; as per the dilemma, such a principle would be either incomplete or intractable. Nor do I think that we can answer the problem of tractability by saying that rationality consists in a specific complexity class (e.g., P), since complexity classes do not describe real-world tractability with sufficient fidelity. Furthermore, although the idea of decision-theoretic rationality generalizes to computational problems beyond betting, the space of algorithms for these problems is complicated and lacks a simple description. Once we have unseated expected-utility maximization as our exact characterization of rationality, no new pretender appears to claim the throne.

Instead, I am for a picture of rationality that looks like the (messy and inexact) picture of real-world computation. It goes something like this: for

[7] The related problem of picking uses for all items to maximize total utility is maximal matching for hypergraphs, which I believe is NP-complete.

any specific problem, such as propositional book evaluation or the TSP, the rational agent should utilize the best available techniques, polynomial and superpolynomial, exact and approximate, and run them for a reasonable amount of time. If his techniques yield an optimal solution before his imprecise or precise-but-arbitrary cutoff, as MMEU does for Elga's Good Book, then rationality requires that he choose it and he is irrational if he does not. But if no such solution materializes, then rationality only constrains him to choose a good approximation.

On this view, the obligation of rationality for any given agent is commensurate with his means. This induces a partial ordering of agents by their ability to be rational; an agent with greater computational power is more obligated than one with less, and so also an agent equipped with the most advanced heuristics over one without them.[8] This is exactly Tsang's "CIDER theory" — "computational intelligence determines effective rationality" — transposed from the descriptive context of behavioral economics to the prescriptive context of philosophy.

But what can we say about rationality simpliciter — unrelativized to the computational ability of any specific agent? My position is that that since the meaning of "tractability" is subject to the Sorites paradox, so is the meaning of "rationality"; the term is irredeemably vague. Three grains of sand are not a heap, and the agent who can't buy the original Good Book or put the gas in the tank is irrational. But 10^{10} grains of sand are a heap, and the agent who can't recognize a Dutch book over hundreds of propositions is not irrational. The obligation of rationality ceases somewhere between these two extremes, but there is no fact of the matter about exactly where it does. On this view, "rational agent" is a notion analogous to the common law's elusive "reasonable man" — when we invoke decision-theoretic rationality we are appealing to an imprecise intuition rather than to something ontologically real.

8 Afterword

Paris quotes significantly more general hardness results than those given here. Here are some relevant ones:

Theorem 3. *The problem of testing the consistency of a set of linear con-*

[8] I think this ordering can only be partial. For example, two agents may possess different heuristics for the same problem, each of which is optimal for a different domain. One example would be a pair of graph algorithms, one of which is faster on sparse graphs, while the other is faster on dense graphs. Another possibility is that one agent has more computational time, i.e., is faster, but the other has more computational space.

straints on personal probability of the form

$$\sum_{j=1}^{r} a_{ji} P(\varphi_{ji}) = b_i$$

where the φ are arbitrary propositional formulae, is NP-complete. (This includes constraints of the form $P(\varphi) = b$ and $P(\varphi \mid \psi) = c$, i.e., unconditional and conditional personal probabilities.)[9]

Theorem 4. *Theorem 3 still holds under any of the following restrictions:*

1. *Allowing only constraints $P(\varphi_i) = b_i$, where the φ_i are 2-ary conjunctions.*

2. *Allowing only constraints $P(p_i) = b_i$ and $P(p_i \mid p_j) = c_i$, where the p_i are atoms.*

The consistency problem is trivial under the following circumstances:

Definition 4. *A* Bayesian network *is a set of probability constraints of the form $P(p_i) = b_i$ and $P(p_i \mid \bigwedge_{j=1}^{r} q_{ij}) = c_i$, where the p_i, q_{ij} are atoms and:*

1. *All atoms are conditionally independent except as specified.*

2. *The directed graph of conditional dependence among atoms is acyclic.*

Every Bayesian network is consistent. But Cooper [1990] proves:

Theorem 5. *Deciding whether $P(p_i) > 0$, for p_i a proposition in an arbitrary Bayesian network, is NP-complete. (However, the value of $P(p_i)$ can be approximated with probabilistic correctness in polynomial time, and if we restrict to networks without* undirected *cycles (i.e., polytrees), we can compute the exact value in polynomial time.)*

When I claimed that the EUM-agent has a tractable algorithm for book evaluation, I allowed the agent an oracle for personal probabilities; I took him at his word, so to speak. But now this is problematic for two reasons:

1. Imagine that an agent is testifying as to his personal probabilities. If they meet even the low bar for complexity described in Theorem 4, I cannot in general determine whether his beliefs are logically possible.

2. Imagine that an agent's personal probabilities form a Bayesian network. Unless this network has a restricted form, the agent cannot in general "know his own mind": he can't compute his exact personal probabilities for propositions.

[9] There is also a positive result here: consistency is in NP, so the complement problem DUTCHBOOK is in co-NP.

These facts seem to pose significant challenges for Bayesian knowledge representation. But they are addressed at a sophisticated level by the engineering literature on graphical models (see Guo and Hsu [2002] for one survey), which fights them by restricting the problem space and moving to approximation algorithms. The picture here seems analogous to our earlier picture of decision-theoretic rationality: clean idealizations are replaced with messy approximations, and philosophy gives way, as it were, to engineering.

On a different note, this paper focuses on the vagueness of Bayesian accounts of rationality in a purely technical context, with respect to a specific well-defined problem. Does this vagueness disappear once we leave the confines of propositional betting books for richer domains, such as science? I think that on the contrary, it becomes much worse. For example, Garber [1983] points out that systematizing all scientific activity within a "global" Bayesian framework would require constructing a universal language of science, in the sense of logical positivism and the Vienna Circle. This would be a striking regression from anti-foundationalism back to foundationalism! In general, claims that Bayesian probability can straightforwardly assimilate our epistemic activity seem to gloss over vital details. I think the purely quantitative vagueness discussed here is only the tip of an iceberg.

9 Acknowledgements

Thanks to Julian Jonker, Sherri Roush, Wes Holliday, Roy Frostig, Justin Vlasits, Adam Lesnikowski, Matt Jones, Umesh Vazirani, Justin Bledin, and Scott Aaronson for helpful discussions.

References

Scott Aaronson. Why philosophers should care about computational complexity. *CoRR*, abs/1108.1791, 2011.

Shabbir Ahmed. Convexity and decomposition of mean-risk stochastic programs. *Mathematical Programming*, 106:433–446, 2006.

Dimitris Bertsimas, Xuan Vinh Doan, Karthik Natarajan, and Chung-Piaw Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Math. Oper. Res.*, 35(3):580–602, 2010.

Andrej Bogdanov and Luca Trevisan. Average-case complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 13(073), 2006.

Cristian Coarfa, Demetrios D. Demopoulos, Alfonso San Miguel Aguirre, Alfonso San, Miguel Aguirre, Devika Subramanian, and Moshe Y. Vardi.

Random 3-sat: The plot thickens. In *In Principles and Practice of Constraint Programming*, pages 143–159, 2000.

Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks (research note). *Artif. Intell.*, 42(2-3):393–405, March 1990. ISSN 0004-3702. doi: 10.1016/0004-3702(90)90060-D. URL `http://dx.doi.org/10.1016/0004-3702(90)90060-D`.

Adam Elga. Subjective probabilities should be sharp. *Philosopher's Imprint*, 10(5), 2010. URL `http://www.princeton.edu/~adame/papers/sharp/elga-subjective-probabilities-should-be-sharp.pdf`.

Daniel Garber. Old evidence and logical omniscience in Bayesian confirmation theory. In *Testing Scientific Theories*, volume X of *Minnesota Studies in the Philosophy of Science*, pages 99–131. University of Minnesota Press, 1983.

Haipeng Guo and William Hsu. A survey of algorithms for real-time Bayesian network inference. In *In the joint AAAI-02/KDD-02/UAI-02 workshop on Real-Time Decision Support and Diagnosis Systems*, 2002. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.5182`.

Ian Hacking. Slightly More Realistic Personal Probability. *Philosophy of Science*, 34(4):311–325, 1967.

Alan Hájek. The reference class problem is your problem too. *Synthese*, 156 (3):563–585, 2007.

Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-sat. *J. Comput. Syst. Sci.*, 62(2):367–375, 2001.

Julian Jonker. Rational decision making with unsharp credences. Circulated manuscript, 2012.

Isaac Levi. *The Covenant of Reason - Rationality and the Commitments of Thought*. Cambridge University Press, 1997. ISBN 978-0-521-57601-7.

Adam Morton. Epistemic virtues, metavirtues, and computational complexity. *Noûs*, 38(3):481–502, 2004. ISSN 1468-0068. doi: 10.1111/j.0029-4624.2004.00479.x.

Robin A. Moser and Dominik Scheder. A full derandomization of schoening's k-sat algorithm. *CoRR*, abs/1008.4067, 2010.

J. B. Paris. *The Uncertain Reasoner's Companion*. Cambridge University Press, Cambridge, UK, 1994.

Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer, 2007.

Edward P. K. Tsang. Computational intelligence determines effective rationality. *International Journal of Automation and Computing*, 5:63–66, 2008. ISSN 1476-8186. doi: 10.1007/s11633-008-0063-6. URL `http://dx.doi.org/10.1007/s11633-008-0063-6`.

Avi Widgerson. Knowledge, Creativity and P versus NP. URL `http://www.math.ias.edu/~avi/PUBLICATIONS/MYPAPERS/AW09/AW09.pdf`. Circulated manuscript, 2009.