HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

Evan Piermont, Peio Zuazo-Garin

FAILURES OF CONTINGENT THINKING

Working Paper WP9/2021/06
Series WP9
Research of economics and finance

Moscow
2021

УДК 330.4
ББК 65.01
Р74

Editor of the Series WP9
“Research of economics and finance”
Maxim Nikitin

Piermont, E., Zuazo-Garin, P.
Р74 Failures of Contingent Thinking [Electronic resource] : Working paper WP9/2021/06 / E. Piermont, P. Zuazo-Garin ; National Research University Higher School of Economics. – Electronic text data (11,3 Mb). – Moscow : Higher School of Economics Publ. House, 2021. – (Series WP9 “Research of economics and finance”). – 43 p.

In this paper, we provide a theoretical framework to analyze an agent who misin- terprets or misperceives the true decision problem she faces. Within this framework, we show that a wide range of behavior observed in experimental settings manifest as failures to perceive implications, in other words, to properly account for the logical relationships between various payoff relevant contingencies. We present behavioral characterizations corresponding to several benchmarks of logical sophistication and show how it is possible to identify which implications the agent fails to perceive. Thus, our framework delivers both a methodology for assessing an agent’s level of contingent thinking and a strategy for identifying her beliefs in the absence full rationality.

УДК 330.4
ББК 65.01

Keywords: Bounded rationality; contingent thinking; subjective beliefs; monotonicity

JEL classification: C72; D81; D84; D91

Piermont Evan, Royal Holloway – University of London, Department of Economics, United Kingdom; E-mail: evan.piermont@rhul.ac.uk

Zuazo-Garin Peio, ICEF, National Research University Higher School of Economics, Russian Federation; E-mail: p.zuazogarin@hse.ru

1 Motivation

When facing complicated and uncertain environments, economic agents often fail to act optimally, choosing actions which do not maximize their expected payoff. An intuitive explanation, and one often appealed to in the experimental and behavioral literature, is that agents misinterpret or misperceive the true decision problem they face—for example, assessing likelihoods in a manner incompatible with classical probability (**??**), failing to properly condition beliefs (**???**), or simply misunderstanding experimental directions or elicitation techniques (**?**).

In this paper, we provide a theoretical framework to analyze an agent whose subjective representation of a decision problem may diverge from the analyst's or experimenter's (objective) representation. Within this framework, we show that a wide range of observed behavior can be explained by agents' failure to *perceive implications*, in other words, to properly account for the logical relationships between various payoff relevant contingencies (this is sometimes referred to as 'failures of contingent thinking'). We present behavioral characterizations corresponding to several benchmarks of rationality—benchmarks employed in the wider theoretical literature—relating the agent's perception of logical implications to her betting behavior, a standard decision theoretic primitive. Our results highlight a fundamental connection between an agent's misperception of implications and her lack probabilistic sophistication, showing that these are often two sides of the same coin.[1] Finally, and perhaps most importantly, we then turn to the identification problem, showing how it is possible within our framework to identify *which* implications the agent fails to perceive.

For example, the contingency "the water is above 150 degrees" implies the contingency "the water is boiling." An agent who would prefer to bet on the former rather than the latter, betrays a belief incompatible with

[1] The relation between contingent thinking and probabilistic reasoning has been previously been explored by **?** and **?**.

understanding this implication. We show that from an agent's preference over bets on contingencies, it is possible to construct a subjective state-space model that entirely characterizes the agent's view of the uncertainty she faces, her level of rationality, and the implications that she perceives.

Almost universally, economic models begin with the proscription of a *state space*, Ω, and model an agent's probabilistic judgements via a probability distribution (or a more general probability-like object), λ, over Ω. Each state $\omega \in \Omega$ is interpreted as the resolution of all uncertainties relevant to the decision problem at hand and λ captures the agent's beliefs about the likelihood of these various contingencies. A state-space, however, exists only insofar as events can be described by an agent. Real interactions with uncertainty, without fail, operate on the level of propositions themselves, rather than on some abstract state-space. Indeed, interpersonal contracts, insurance policies, economics experiments, etc. all describe the resolution of uncertainty by directly representing the contingencies in human language. While it is well known that under sufficient logical omniscience or rationality the abstraction to a state space is without loss of generality, this is not the case for agents who are not perfect contingent reasoners; in particular, assumptions about the structure of the state space impose tacit restrictions on how the agent interprets contingencies and thus how she might fail at contingent thinking.

Therefore, our paper begins with the more primitive notion of eliciting an agent's uncertainty about verbal or linguistic contingencies (like "the water is boiling") rather than uncertainty about states in an exogenous state space. The first set of results then shows how it is possible to *construct* a state space (and probability) that faithfully represents the agent's uncertainty. In other words, the constructed state space represents the agent's interpretation of the contingencies, and hence, her ability to recognize logical relationships such as implication. As such, the interpretation of the contingencies is a subjective aspect of the representation, rather than an exogenously (and implicitly) imposed condition. These results provide a direct method of testing the rationality of economic agents, and for identifying the agent's perception regrading logical implications between contingencies.

Some implications, such as the relation between "the temperature is above 150 degrees" and "the water is boiling," rely on a physical background theory. An agent's failure to perceive this implication might arise from her not considering the same theory as the modeler (perhaps she interprets 'degrees' in Fahrenheit rather than Celsius or considers a different ambient pressure). Thus, an agent might take actions inconsistent with the analyst's predictions either because (i) she fails some criterion of rationality or (ii) she is rational but operating under a different background theory.

While our above mentioned behavioral characterizations allows the analyst to discern between these two types of agents, our next set of results take on the task of identifying, in the latter case, the subjective theory of the agent. For example, an experimenter could find that a subject who is taking suboptimal actions may be neither irrational nor have non-standard preferences, but instead have misinterpreted the experimental instructions. In other words, the subject is acting optimally conditional on her misinterpretation. In such a state of affairs, we provide the experimenter with the tools to identify which of the instructions the subject failed to understand. Thus, our framework delivers a methodology for the analyst to both asses the agent's level of contingent thinking and identify to her beliefs even in the absence of the usual tenets of rationality.

The reminder of the paper is structured as follows: The next section outlines our model and main results by way of two extended examples. These examples are taken directly from the experimental literature and exhibit how our model can be used to explain real word decision making. Then, Section 3 introduces the model and primitive. Section 4 then provides some basic representation results relating the DM's observable behavior to her ability to preform contingent thinking. Section 5 contains our identification results, exploring how our model allows a modeler to identify the theory of an agent. A discussion of related literature is contained in Section 6. Appendix A contains additional representation results. The Appendix B is an application of the model to games: we show how our model can be used to construct a Wald-Pearce type criterion determining rationalizability under non-expected utility beliefs. Finally, Appendix C contains proofs omitted from the main text.

2 An Overview of our Syntactic Model and Results

We take as given a set of (logically connected) statements, $\mathcal{L}$. These statements regard the uncertainties faced in a decision problem: for example "the temperature is above 150 degrees" or "the pressure is 1 bar and the temperature is above 150 degrees." We assume the modeler can observe the agent's probabilistic assessment of statements, $\{\pi(\varphi) \in [0,1] \mid \varphi \in \mathcal{L}\}$, where $\pi(\varphi)$ is interpreted as the probability the agent places on φ being true.

Since the state-space is not exogenously given, we are interested in *subjective models of uncertainty.* A subjective model of uncertainty is a triple (Ω, t, λ):

- Ω, a *state space*, is a set.
- $t : \mathcal{L} \to 2^{\Omega}$, a *truth valuation*, maps each statement to a subset of states interpreted as those states in which the statement is true.
- $\lambda : 2^{\Omega} \to [0,1]$, a *likelihood assessment*, maps subsets of the state space to the 0-1 interval (with $\lambda(\varnothing) = 0$ and $\lambda(\Omega) = 1$), interpreted as the likelihood of the event.

Therefore, an agent whose view of uncertainty is represented by (Ω, t, λ) views the states in Ω as the possible resolutions of uncertainty regarding the statements of $\mathcal{L}$: at ω exactly the statements $\{\varphi \mid \omega \in t(\varphi)\}$ are perceived to be true. Then λ determines the the agent's belief about the likelihood of these various resolutions. In the usual way—for example, by relaxing additivity—λ can accommodate failures in probabilistic reasoning.

Similarly, t can accommodate failures in logical reasoning, if for example t(“the pressure is 1 bar and the temperature is above 150 degrees”) $\neq t$(“the pressure is 1 bar”)$\cap$ t(“the temperature is above 150 degrees”).

As a first task, we ask when is it possible to find a subjective model of uncertainty, (Ω, t, λ), that faithfully captures the agent’s observable beliefs: that is, for each $\varphi \in \mathcal{L}$, the likelihood that φ is true as given by the model (Ω, t, λ) is the same as the observed assessment $\pi(\varphi)$: in math, $\lambda(t(\varphi)) = \pi(\varphi)$. We show that it is in general possible to construct a state-space representation, and that different assumptions on rationality impose different constraints on the truth valuation and likelihood assessment. This is important as is allows the modeler to work with the more familiar state-space representation, both while using a primitive that is more realistically observable and while allowing for errors in reasoning.

We ask further how the properties of t and the properties of λ relate and more generally, how these properties are related to conditions on the observable. Since t captures the agent’s logical reasoning and λ her probabilistic reasoning, our model allows for the distinction between these two separate departures from from the rational benchmark. We show that there is a tradeoff between these two forms of reasoning, as evidenced by the following example, and explained in detail afterwards.

Example 1. **?** provided subjects with the following vignette:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Linda is a teacher in elementary school. Linda works in a bookstore and takes Yoga classes.

and asked them rank statements in order of likelihood including the following three:

F = “Linda is active in the feminist movement.”

$\textsc{t}$ = “Linda is a bank teller.”

$\textsc{t} \wedge \textsc{f}$ = “Linda is a bank teller and is active in the feminist movement.”

85% of subjects ranked $\textsc{f} > \textsc{t} \wedge \textsc{f} > \textsc{t}$ in opposition to any objective (i.e., sound) model of uncertainty, which on the basis that $\textsc{t} \wedge \textsc{f}$ implies $\textsc{t}$, must rank the later weakly more likely.

To make matters more concrete, lets assume that a subject assess the likelihoods as $\pi(\textsc{f}) = \frac{3}{4}$, $\pi(\textsc{t} \wedge \textsc{f}) = \frac{1}{2}$, and $\pi(\textsc{t}) = \frac{1}{4}$. First, notice that we can represent these assessments via a subjective model of uncertainty (Ω, t, λ) with a probability measure λ, by permitting flexibility in t: set $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and $t : \textsc{f} \mapsto \{\omega_1, \omega_2\}, \textsc{t} \mapsto \{\omega_2\}$ and $\textsc{t} \wedge \textsc{f} \mapsto \{\omega_2, \omega_3\}$. Let λ be the probability measure given by $\lambda(\omega_1) = \frac{1}{2}$ and $\lambda(\omega_2) = \lambda(\omega_3) = \frac{1}{4}$. Clearly, $\lambda(t(\varphi)) = \pi(\varphi)$ for each statement φ, but representing the subject's departure from rationality forces the truth valuation to be logically flawed: despite $\textsc{t}$ being a logical consequence of $\textsc{t} \wedge \textsc{f}$, in state ω_3, the later is true but the former is not.

From this, one might be tempted to conclude that the subject has perfect probabilistic judgment (setting the interpretation of statements aside, her probabilistic reasoning as given by λ is additive and so coincides with the rational benchmark). However, (Ω, t, λ) is but one of many representations of her observable behavior:

Consider (Ω, t', λ') with $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $t' : \textsc{f} \mapsto \{\omega_1, \omega_2\}, \textsc{t} \mapsto \{\omega_2, \omega_3\}$, and as dictated by the logical connection between statements $\textsc{t} \wedge \textsc{f} \mapsto \{\omega_2\}$. Finally, set $\lambda'(\{\omega_i\}) = \frac{1}{2}$ for $i \in \{1, 2, 3\}$, $\lambda'(\{\omega_1, \omega_2\}) = \lambda'(\{\omega_1, \omega_3\}) = \frac{3}{4}$, and $\lambda'(\{\omega_2, \omega_3\}) = \frac{1}{4}$. Again, we have a representation of the subject's assessments, but this time, the truth valuation respects all logical implications. Of course, λ is *not* a probability distribution; far from it, it is not even monotone: the failure to perceive the implication $\textsc{t} \wedge \textsc{f} \Rightarrow \textsc{t}$ is reflected in the fact that $\lambda(t(\textsc{t})) = \lambda(\{\omega_2, \omega_3\}) = \frac{1}{4} < \frac{1}{2} = \lambda(\{\omega_2\}) = \lambda(t(\textsc{t} \wedge \textsc{f}))$. □

In Example 1, while it is clear that subject fails to perceive the implication $\textsc{t} \wedge \textsc{f} \Rightarrow \textsc{t}$, the multiplicity of representations indicates that this misperception can be captured as either a failure in logical or probabilistic reasoning. In a series of duality results, we show that this is universally the case: for a large class of behavior whatever can be explained by failures of probabilistic reasoning alone can be likewise rationalized by failures of contingent thinking alone, and vice versa.[2]

While the entanglement of probabilistic and logical reasoning may initially feel disappointing, we argue that this collapse is both philosophically sensible and practically beneficial. On the pragmatic side, deferring the discussion of abstract motivations, it is this duality that allows the modeler to leverage observable betting behavior—a standard decision theoretic primitive and one often associated squarely with probabilistic judgements—to identify much about the agent's cognition under limited contingent reasoning. The subject's failure to realize $\text{T} \wedge \text{F} \Rightarrow \text{T}$, is unambiguously captured by the agents preference to bet on $\text{T} \wedge \text{F}$ rather than T. By appealing to this observable primitive, we define what it means for an agent to fail to perceive an implication while remaining agnostic as to whether these failures are modeled directly via logical relationships (i.e., non-rational t) or indirectly via relaxations of probabilistic thinking (i.e., non-rational λ). Using such a primitive, we show it is easy to identify the set of logical implications perceived by the agent.

Often the implications in question are the result of a background theory (for example, that water boils at 100 degrees). An agent's failure to perceive an implication predicated on the modeler's background theory could be a consequence of flawed logical reasoning or of the agent entertaining a different theory. Our next set of results show that it is possible to identify the largest sub-theory (of the modeler's theory) that the agent understands.

Example 2. **?** provided subjects with the following decision problem (paraphrased):

[2] However, outside of this class of behavior, we show that failures in contingent thinking are the strictly more general explanation.

> A *selected ball* is chosen to be red with probability $\alpha \in (0, 1)$ and otherwise blue. The subject must cast a vote for either RED or BLUE without observing the color of the selected ball. In addition, two computers observe the color of the selected ball and are programmed to follow specific rules for casting a vote in favor of RED or BLUE that are contingent on the color of the selected ball. If the color chosen by a simple majority matches the color of the selected ball, the subject's payoff is \$2; otherwise, the payoff is \$0. Before casting her vote, the subject receives information about the rule being followed by the computers and the probability p, but does receive information about the actual votes of the computers. Both computers follow the rule: (i) If selected ball is red: vote RED; (ii) If selected ball is blue: vote BLUE with probability β and RED with $1-\beta$ for fixed (and known) $\beta \in (0, 1)$.

Notice that voting BLUE is a (weakly) dominant strategy since in order for a subject's vote affect the outcome, the computers must disagree, and hence the selected ball must be blue. Nevertheless, 80% of subjects do not play strategically even after 40 rounds of play.[3]

Let us consider the following statements (and their disjunctions, negations, etc.) regarding the decision problem:

$\textsc{r} =$ "The selected ball is red."

$\textsc{b} =$ "The selected ball is blue."

$\textsc{p} =$ "The subject's vote will determine the outcome."

The (relevant for the discussion) rules of the game can be encoded by a modelers theory: $\mathcal{T} = \{\textsc{r} \leftrightarrow \neg\textsc{b}, \textsc{p} \rightarrow \textsc{b}\}$—which can be read as "red if and only if not blue," and "if the subject is pivotal then blue."

Assume for the sake of argument that the modeler elicited the subject's assessment of the likelihood of the various statements and found

[3] This was defined as voting RED more than 15% of the time.

1. $\pi(\textsc{r}) = \pi(\neg\textsc{b}) = 1 - \pi(\textsc{b}) = 1 - \pi(\neg\textsc{r}) = \alpha$,
2. $\pi(\textsc{p}) > 0$, and,
3. $\pi(\textsc{p} \wedge \textsc{r}) \geq \pi(\textsc{p} \wedge \textsc{b})$

(1) states the subject understands the way the ball is drawn, (2) states the subject considers it possible she is pivotal and (3) corresponds to the preference that, when her vote will determine the outcome, she prefers to vote RED.

The subject's behavior cannot be explained by any model in which she understands all implications of the background theory $\mathcal{T}$. Indeed, say the subject represents these statements via a subjective model of uncertainty (Ω, t, λ). Notice that under $\mathcal{T}$, $\textsc{p}$ is logically equivalent to $\textsc{p} \wedge \textsc{b}$ and $\textsc{r} \wedge \textsc{p}$ is a contradiction (i.e., is logically equivalent to $\mathbf{F}$, the variable standing for false). So if the subject understands these implications we have $t(\textsc{r} \wedge \textsc{p}) = \varnothing$ and $t(\textsc{p}) = t(\textsc{b} \wedge \textsc{p})$. But this means $\pi(\textsc{r} \wedge \textsc{p}) = \lambda(t(\textsc{r} \wedge \textsc{p})) = 0 < \pi(\textsc{p}) = \lambda(t(\textsc{p})) = \lambda(t(\textsc{b} \wedge \textsc{p})) = \pi(\textsc{b} \wedge \textsc{p})$, violating (3).

Now, consider $\Omega = \{r, b\} \times \{p, \neg p\}$ and let t send propositions to the relevant cylinder sets (i.e., $t(\textsc{r}) = \{(r, p), (r, \neg p)\}$, $t(\textsc{r} \wedge \textsc{p}) = (r, p)$, $t(\textsc{r} \vee \textsc{b}) = \Omega$, etc) . Let λ be a measure such that $\lambda(\{(r, p), (r, \neg p)\}) = \alpha$ and $\lambda((r, p)) > \lambda((b, p))$. Obviously, such a mapping captures the likelihood assessments (1)-(3). Also, it is completely rational (the mapping t obeys all logcial necessities, and λ is additive) despite the fact that it ignores the connections between $\textsc{b}$ and $\textsc{p}$. The agent using this model is rational but does not consider the theory $\mathcal{T}$—although she does consider (and hence perceives all implications of) the sub-theory $\{\textsc{r} \leftrightarrow \neg\textsc{b}\} \subset \mathcal{T}$. □

Our results below provide a test for whether the agent is logically flawed or possibly just entertaining a different theory. In the later case, we show that this theory is identifiable, and in fact, easily constructed from the observable primitive. Notice that in the model at the end of Example 2, the statement $\textsc{r} \leftrightarrow \neg\textsc{b}$ was true at every state: if an agent perceives all implications of a theory $\mathcal{T}$ then she assigns probability 1 to the statements that compose $\mathcal{T}$.

We further show that under slightly stronger rationality assumptions, this converse of this relationship holds also, so that we can identify the theory understood by the agent by collecting the statements to which she assigns probability 1.

3 Framework

3.1 Preliminaries

Let $\mathbb{P}$ collect a set of propositional variables: statements about the world that can either be true of false. For example “It is raining” or “The S&P500 went up today.” We assume that $\mathbb{P}$ contains two distinguished propositions $\mathbf{T}$ and $\mathbf{F}$, that are interpreted as “true” and “false” respectively. Then, $\mathcal{L}(\mathbb{P})$ is the language defined inductively, beginning with $\mathbb{P}$ and such that if φ, ψ are in $\mathcal{L}(\mathbb{P})$ then so too are $\neg\varphi$, $(\varphi \wedge \psi)$, and $(\varphi \vee \psi)$. The interpretation is as in propositional logic: $\neg\varphi$, the *negation* of φ, is interpreted as the statement that φ is not true, $\varphi \wedge \psi$, the *conjunction* of φ and ψ, is interpreted as the statement that both φ and ψ are true, and $\varphi \vee \psi$, the *disjunction* of φ and ψ, is interpreted as the statement that at least one of φ and ψ is true.

Write $\varphi \Rightarrow \psi$, if ψ can be deduced from φ under the deduction rules (and logical tautologies) of propositional logic. For example, $\textsc{p} \Rightarrow \textsc{p} \vee \textsc{q}$ is valid implication. Take as short-hand $\varphi \Leftrightarrow \psi$ to mean that $\varphi \Rightarrow \psi$ and $\psi \Rightarrow \varphi$, i.e., that ψ and φ are logically equivalent. It is well known that $\Leftrightarrow$ is an equivalence relation, so let $\mathbf{LT}(\mathcal{L})$ collect the equivalence classes of logically equivalent formula. $\mathbf{LT}(\mathcal{L})$ is called the Lindenbaum–Tarski algebra of $\mathcal{L}$, and it is a fact that $\mathbf{LT}(\mathcal{L})$ is a Boolean Algebra under the operations inherited from the grammar.

If Ω is a set, then call $t : \mathcal{L} \to 2^{\Omega}$ a *truth valuation* if $t(\mathbf{T}) = \Omega$ and $t(\mathbf{F}) = \varnothing$. The interpretation is that $t(\varphi)$ is the set of states where φ is true. Call t

- *exact* if $\varphi \Leftrightarrow \psi$ implies $t(\varphi) = t(\psi)$.

- *monotone* if $\varphi \Rightarrow \psi$ implies $t(\varphi) \subseteq t(\psi)$.
- *symmetric* if $t(\neg\varphi) = \Omega - t(\varphi)$
- $\wedge$-*distributive* if $t(\varphi \wedge \psi) = t(\varphi) \cap t(\psi)$
- *logically sound* (or just *sound*) if it is all of the above

Remark 1. It follows from the well known fact that Boolean algebras provide semantics for propositional logic—see for example **?**—that $\psi \Rightarrow \varphi$ iff for every state-space Ω and every sound $t : \mathcal{L} \to \Omega$ we have $t(\psi) \subseteq t(\varphi)$.

For (Ω, Σ), with Ω a set and $\Sigma \subseteq 2^{\Omega}$ a field of sets, call a map $\lambda : \Sigma \to [0,1]$ a *likelihood appraisal* if $\lambda(\varnothing) = 0$ and $\lambda(\Omega) = 1$. We denote the set of likelihood appraisals for (Ω, Σ) by $\Lambda(\Omega, \Sigma)$. Likelihood appraisals are referred to as different things in different literatures: normalized games, grounded normalized set functions, non monotone capacities, etc.

Call $\lambda \in \Lambda(\Omega, \Sigma)$,

- *symmetric* if for all $A \in \Sigma$, $\lambda(A) = 1 - \lambda(A)$
- *monotone* (or a *capacity*) if for all $A, B \in \Sigma$ with $A \subseteq B$, $\lambda(A) \leq \lambda(B)$
- *totally monotone* if for any $A_1 \ldots A_n \in \Sigma$,

$$\lambda\Big(\bigcup_{i \leq n} A_i\Big) \geq \sum_{\{I \mid \varnothing \neq I \subseteq \{1 \ldots n\}\}} (-1)^{|I|+1} \, \lambda\Big(\bigcap_{i \in I} A_i\Big),$$

- *additive* (or a *measure*) if for all $A, B \in \Sigma$ with $A \cap B = \varnothing$, $\lambda(A \cup B) = \lambda(A) + \lambda(B)$.

Remark 2. Total monotonicity implies monotonicity, and additivity is equivalent to total monotonicity and symmetry, see for example, **?** Figure 2.1.

Likelihood appraisals carry well-defined theory of integration: for each measurable $x : \Omega \to \mathbb{R}$ and likelihood appraisal λ let,

$$\int x \mathrm{d}\lambda = \int_{-\infty}^{\infty} \lambda(\{\omega \in \Omega \mid x(\omega) \geq r\}) \mathrm{d}r. \tag{1}$$

where the integral on the right hand side is the standard Lebesgue integral over $\mathbb{R}$. Notice that (1) is additive over pairwise co-monotone functions. When λ is a capacity, (1) coincides with the usual definitions of a Choquet integration, and when it is additive, (1) becomes the usual additive integral.

Given $\mathcal{L}$ a triple (Ω, t, λ) is called a *subjective model of uncertainty* for $\mathcal{L}$ where $t : \mathcal{L} \to 2^{\Omega}$ is a truth valuation and $\lambda \in \Lambda(\Omega, \Sigma)$ with $t(\mathcal{L}) \subseteq \Sigma$.

3.2 Primitive

The primitive of our model is a preference relation over *bets*. A *primitive bet on* φ is a function $b_{\varphi} : \mathcal{L} \to [0, 1]$ such that $b(\varphi) = 1$ and $b(\psi) = 0$ for all $\psi \neq \varphi$. The interpretation of b_{φ} is a bet that pays 1 (util) when φ is true and is called off otherwise. A (general) *bet* is a finitely supported lottery over primitive bets. That is a bet is of the form $b : \mathcal{L} \to [0, 1]$ such that $\text{supp}(b) = \{\varphi \in L \mid b(\varphi) > 0\}$ is finite, and with $\sum_{\varphi \in \text{supp}(b)} b(\varphi) = 1$. The interpretation of $b(\varphi)$ is the likelihood of receiving the primitive bet on φ. We sometimes write $\{\alpha_i b_{\varphi_i} \ldots \alpha_n b_{\varphi_n}\}$ to denote the b such that $\text{supp}(b) = \{\varphi_i \ldots \varphi_n\}$ and $b(\varphi_i) = \alpha_i$ (where, obviously, the α's are positive and sum to 1).

The set of all bets, $\mathcal{B}$, is a mixture space when mixtures are taken pointwise. Note, however, this is not the usual set of acts from an **?** type framework—for example, $\mathcal{B}$ does not contain constant acts—as we have inverted the order of subjective and objective uncertainty. The domain of bet, $\mathcal{L}$, is not a state-space, but a set of statements representing uncertain propositions. The state space, which encodes the possible resolutions of this uncertainty, is part of the representation rather than the primitive.

We assume that $\succcurlyeq$ entertains an expected utility structure, which we axiomatize explicitly below. This may initially seem odd, as we are interested in detailing failures of rationality which include failures of probabilistic reasoning. However, our language based framework explicitly separates bets (on linguistic constructions based on $\mathcal{L}$) and acts as functions from a state-space to utils (that will be part of the representation). Thus, our assumption of additivity is emphatically not regarding the decision maker's beliefs, but rather her evaluation of the objective uncertainty over primitive bets. We simply are assuming that if the decision maker values a bet of φ at $\frac{1}{2}$ and values a bet of $\neg\varphi$ at $\frac{1}{4}$ then she evaluates the half-half mixture between them at $\frac{3}{8}$. It is, at the same time, permissible that she evaluates a bet on $\varphi \vee \neg\varphi$ at 1 (or any other value).

A subjective model of uncertainty (Ω, t, λ) *represents* $\succcurlyeq$ if, for all $b, b' \in \mathcal{B}$, $b \succcurlyeq b'$ if and only if

$$\sum_{\varphi \in \text{supp}(b)} b(\varphi)\lambda(t(\varphi)) \geq \sum_{\varphi \in \text{supp}(b')} b'(\varphi)\lambda(t(\varphi)).$$

4 Representations

4.1 Axioms

If we do not care at all about the properties of t then it is always possible to construct a state space that represents $\succcurlyeq$, so long as these preferences obey some basic conditions relating to the management of objective risk:

Axiom 1—Non-Triviality (**NT**). $b_{\mathbf{T}} \succcurlyeq b_{\varphi} \succcurlyeq b_{\mathbf{F}}$ for all $\varphi \in L$ and $b_{\mathbf{T}} \succ b_{\mathbf{F}}$.

Axiom 2—Objective Expected Utility (**EU**). $\succcurlyeq$ is a complete, transitive and satisfies the Archimedean and Independence axioms.

Axiom **NT** states two things: first that the preferences are non-trivial, and second that the agent understands what **T** and **F** represent. This latter restriction is tantamount to assuming that "receive x no matter what" and "receive x never" are unambiguous. Axiom **EU** plays the obvious role of providing a cardinal measure between lotteries—as discussed above, it deals only with how the agent perceives objective risk and not the resolution of statements in $\mathcal{L}$.

Proposition 1. *The following are equivalent:*

1. $\succcurlyeq$ *satisfies Axioms* **NT** *and* **EU**, *and*

2. $\succcurlyeq$ *is represented by a triple* (Ω, t, λ)*, with* λ *additive.*

The state-space constructed in the proof of Proposition 1 is extraordinarily wasteful. Essentially, in constructs independent state spaces for each statement–exploiting the fact that t can be completely unrestricted—so that $t(\varphi) \neq t(\psi)$ for *any* statements, despite the agent perhaps understanding some logical equivalences.

In what follows we introduce additional axioms on $\succcurlyeq$, nested by successive strength, that allow the construction of subjective models with ever more structure. Of course, as we introduce more structure on t or λ, we loose the ability to represent certain failures of contingent reasoning and the constructed state spaces look closer and closer to those of pure rationality. Ultimately we wind up with the prototypical rational model used in economics, with a sound t and additive λ.

The first rationality axiom concludes that the agent understands logical equivalence, even if she misunderstands the relationship between other statements.

Axiom 3—Equivalence (**E**). If $\psi \Leftrightarrow \varphi$ then $b_\psi \sim b_\varphi$.

As the next proposition makes evident, Axiom **E** is a necessary condition to model behavior with a sound t. Thus, no matter how much we insist on capturing failures of reasoning via relaxations of probabilistic sophistication, without imposing Axiom **E** we cannot find a representation.

Proposition 2. *The following are equivalent:*

1. $\succcurlyeq$ *satisfies Axioms* **NT**, **EU** *and* **E**,

2. $\succcurlyeq$ *is represented by a triple* (Ω, t, λ)*, with* t *exact and* λ *additive, and,*

3. $\succcurlyeq$ *is represented by a triple* (Ω', t', λ')*, with* t' *sound.*

Primitive	Representation	
NT, **EU** and	λ additive and	t sound and λ:
$\varnothing$	any t	N/A
E	t exact	any λ
I	t monotone	λ monotone
IE	t $\wedge$-distributive	λ totally monotone
A	t sound	λ additive

Figure 1: The relationship between axioms and representations. The last two equivalences are provided in Appendix A.

Next, we assume that if the truth of ψ follows logically from the truth of φ then the agent would prefer a bet on ψ than on φ. The interpretation, which will be substantiated not only by the immediate proposition but also by Proposition 4, is that the agent recognizes and understands implications. If $\varphi \Rightarrow \psi$ then whenever φ is true, so too is ψ, but it might still be that ψ alone is true, so, an agent cognizant of this should prefer to bet on ψ.

Axiom 4—Implication (**I**). If $\varphi \Rightarrow \psi$ then $b_\psi \succcurlyeq b_\varphi$.

The content of Axiom **I** is that the resulting representation are monotone. Thus, under the interpretation that mapping the antecedent φ to a subset of consequent ψ is *understanding* the implication $\varphi \Rightarrow \psi$, then assuming an agent's probabilistic judgements over a state space are given by a capacity is implicitly assuming that the agent understands all implications.

Proposition 3. *The following are equivalent:*

1. $\succcurlyeq$ *satisfies Axioms* **NT**, **EU**, *and* **I**,
2. $\succcurlyeq$ *is represented by a triple* (Ω, t, λ), *with* λ *additive and* t *monotone, and,*
3. $\succcurlyeq$ *is represented by a triple* (Ω', t', λ'), *with* λ' *monotone and* t' *sound.*

In Appendix A, we provide two additional representation results. First, **IE** strengthens **I** to deal with the case when multiple statements all imply the same consequent. By placing bounds on the type of non-additivity that can enter the model, the ensuing preference is represented by a totally monotone capacity or a $\wedge$-distributive truth function. Second, **A** further strengthens **IE** so that the likelihood of the disjunction of pair of mutually incompatible statements is the sum of their individual likelihoods. This ensures ‘full’ rationality, the agents preferences are representable by a model with both a sound t and an additive λ. This spectrum of duality results as induced by conditions on the primitive is summarized in Figure 1.

4.2 Identification

When observing an agent’s choices in a decision problem, a modeler will want to make inference about the agent’s reasoning capabilities. However, the representation theorems in Section 4 show that in general there will not be a unique subjective model that represents the agent’s preferences. This is problematic, as in some models representing the agents preferences, it might be that $t(\varphi) \subseteq t(\psi)$ while in others $t'(\varphi) \not\subseteq t'(\psi)$, leaving the matter of the agent’s belief in the implication $\varphi \Rightarrow \psi$ unresolved. As such we would like a notion of inference that deals directly with primitives and is therefore invariant in the choice of representation.

For two statements, $\varphi, \psi \in \mathcal{L}$ such that $\varphi \Rightarrow \psi$, say that the agent (given by $\succcurlyeq$) *understands that* φ *implies* ψ if there exists a subjective model of uncertainty, (Ω, t, λ), with λ additive and such that $t(\varphi) \subseteq t(\psi)$. This is the most generous assessment of the agent’s understanding—we assume that the agent understood an implication unless it is impossible to find a model in which the implication is respected.[4] The most conservative assessment, by contrast, where we define understanding to be if and only if $t(\varphi) \subseteq t(\psi)$ *for all* models, is trivial: the agent would understand nothing except that

[4] It may seem contrived that we look only at representations with additive λ, but it is essentially without loss of generality. Indeed, the more permissive definition, which requires only that $t(\varphi) \subseteq t(\psi)$ and $\lambda(t(\varphi)) \leq \lambda(t(\psi))$ leads to the same characterization. Additionally, by considering additive likelihood assessments, we narrow in failures of logical thinking.

$\mathbf{F} \Rightarrow \varphi \Rightarrow \mathbf{T}$ (this is a consequence of the construction in the proof of Proposition 1). Indeed, it is always possible to construct a model in which the agent understands no non-trivial implications at all, but where her preferences just happen workout as if she did.

PROPOSITION 4. *Let $\varphi, \psi \in \mathcal{L}$ be such that $\varphi \Rightarrow \psi$. Then an agent (given by $\succcurlyeq$ satisfying* **NT** *and* **EU***) understands that φ implies ψ if and only if $b_\psi \succcurlyeq b_\varphi$.*

Thus, our notion of understanding, which required quantifying over all possible representations, is in fact captured directly by the primitive. The interpretation of this result is far deeper than its technically simple proof might indicate: a modeler cannot conclude with certainty that an agent fails to understand the implication $\varphi \Rightarrow \psi$ unless the agent would prefer to bet on φ than on ψ. Any indirect method of assessment, through the use of more complex compound statements, or whatever else, must eventually reduce to perceiving an antecedent as more likely than its consequent. This result also shows in full force the duality between probabilistic judgments and contingent thinking. Failures to perceive implications, as permitted by the construction of a model with a flawed t, are exactly non-monotonicities in probabilistic judgement.

5 Identifying Theories

All of the above deals only with understanding implications that are dictated purely by the rules of logic and so do not depend in any way on the interpretation of the propositional variables. That $\textsc{p} \wedge \textsc{q} \Rightarrow \textsc{p}$ in no way depends on the meaning of the statements $\textsc{p}$ and $\textsc{q}$. Often, however, the implications we have in mind arise from non-logical relationships between propositional variables. For example, if $\textsc{p} =$ "the water is boiling" and $\textsc{q} =$ "the temperature is above 150 degrees" then our *physical* understanding of the world dictates that $\textsc{q}$ implies $\textsc{p}$, despite them being *logically* independent statements.

The relationship between propositional variables can be encoded by a collection of statements, called a theory, that are presupposed to be true. This physical theory of water above can be captured by the presupposition of $\mathcal{T} = \{\neg\text{Q} \vee \text{P}\}$. Given the the theory $\mathcal{T}$, P can be deduced from Q (i.e., P can be deduced from Q *and* the elements of the theory, under the usual rules of propositional logic). In order for a theory to be sensible it must be consistent, that is, it must not contain statements that contradict one another.

A modeler will often have in mind some specific theory, $\mathcal{T}$, and will judge the reasoning capabilities of agents not against logical necessity but also their understanding of $\mathcal{T}$. An agent who believes it more likely the water is above 150 degrees than that it is boiling, exhibited by $b_{\text{Q}} \succ b_{\text{P}}$, might be a flawed logical reasoner, or she might misunderstand the meaning of the statements (perhaps interpreting 'degrees' in Fahrenheit rather than Celsius). In what follows we provide a framework for disentangling these two notions, and for identifying the largest sub-theory of the modelers theory that the agent understands.

Formally, call $\mathcal{T} \subset \mathcal{L}$ a *theory* if (i) it is closed under logical implication, and (ii) $\mathbf{F} \notin \mathcal{T}$.

Remark 3. It follows from the compactness theorem for propositional logic—see for example **?**—that if no contradiction can be derived from any finite subset of $\mathcal{S} \subset \mathcal{L}$ then $\mathcal{S}$ is satisfiable in its entirety. Thus, if $\mathcal{S} \subset \mathcal{L}$ is such that there does not exist a finite set $\mathcal{S}' \subseteq \mathcal{S}$ such that $\bigwedge_{\mathcal{S}'} \varphi \Rightarrow \mathbf{F}$, then the closure of $\mathcal{S}$ under implication, $\mathcal{T}$, is a theory and there exists a (Ω, t) with t sound such that $t(\varphi) = \Omega$ for all $\varphi \in \mathcal{T}$. It is therefore without loss of generality to identify a non contradictory collection of statements with the theory induced by closing off under implication.

Given a theory $\mathcal{T}$, we can consider the theory specific notion of implication: $\varphi \overset{\mathcal{T}}{\Rightarrow} \psi$ if ψ can be deduced from φ and $\mathcal{T}$ under the deduction rules (and logical tautologies) of propositional logic.[5]

For example, if $\mathcal{T} = \{\neg\text{Q} \vee \text{P}\}$, $\text{Q} \overset{\mathcal{T}}{\Rightarrow} \text{P}$, although it is not true that $\text{Q} \Rightarrow \text{P}$. Therefore, the distinction between $\Rightarrow$ and $\overset{\mathcal{T}}{\Rightarrow}$ provides the distinction between being a flawed logical reasoner (failing to understand a $\Rightarrow$ implication) and not entertaining the same theory (failing to understand a $\overset{\mathcal{T}}{\Rightarrow}$ implication), where failing to understand an implication of either type means preferring to bet on the antecedent (i.e., generalizing the definition in Section 4.2).

That is to say: we can repeat the exercise presented in Section 4 using $\overset{\mathcal{T}}{\Rightarrow}$ in place of vanilla implication. Specifically, say that t is $\mathcal{T}$*-monotone* if $\varphi \overset{\mathcal{T}}{\Rightarrow} \psi$ then $t(\varphi) \subseteq t(\psi)$, and likewise say that $\succcurlyeq$ satisfies $\mathcal{T}$-**I** if $\varphi \overset{\mathcal{T}}{\Rightarrow} \psi$ then $b_\psi \succcurlyeq b_\varphi$. It is straightforward to see that the relation established in Proposition 3 continues in this $\mathcal{T}$-specific setup. That is, $\succcurlyeq$ satisfies $\mathcal{T}$-**I** if and only if it can be represented by some (Ω, t, λ) with an additive λ and a $\mathcal{T}$-monotone t.

This generalization does much more that simply allow us to discuss theory specific implication: it allows a modeler to *identify* the theory that underlies the agent's preferences. Suppose the modelers conception of the world was given by a theory $\mathcal{T}$.[6] If an agent understands all logical implications of her knowledge, but considers a different theory than the modeler then her preferences will satisfy **I** but not $\mathcal{T}$-**I**.

It is of practical value then to try and elicit what is the largest sub-theory of $\mathcal{T}$ consistent with the agent's preferences. The following result states that this is possible: there exists a unique largest sub-theory of $\mathcal{S} \subseteq \mathcal{T}$ such that the agent understands all $\mathcal{S}$-implications (i.e., the agent fails to understand some implication of any proper superset of $\mathcal{S}$).

[5] Thus we could view regular implication as the theory specific implication induced by the set of all logical tautologies.

[6] If we insist that the modeler harbors no uncertainty about the decision problem—when, for example, the analysis is done ex-post—then $\mathcal{T}$ will be maximal: it will contain either φ or $\neg\varphi$ for all $\varphi \in \mathcal{L}$. Any representation of a maximal theory has $\lambda(t(\varphi)) \in \{0, 1\}$, so all uncertainty is resolved. In fact the converse is also true: the set of maximal theories corresponds exactly the set non-trivial $\{0, 1\}$-valued finitely additive measures over $\mathbf{LT}(\mathcal{L})$.

Proposition 5. *Let $\succcurlyeq$ satisfy* **NT**, **EU** *and* **I**. *Let $\mathcal{T}$ be a theory. Then there exists a unique sub-theory $\mathcal{S} \subseteq \mathcal{T}$ such that $\succcurlyeq$ satisfies $\mathcal{S}$-***I** *and for any $\mathcal{S} \subset \mathcal{S}' \subseteq \mathcal{T}$, $\succcurlyeq$ does not satisfy $\mathcal{S}'$-***I***.*

Thinking along the lines of Example 2, the rules of a economics experiment can be coded in a theory $\mathcal{T}$ which is presupposed by the modeler. A subject who makes choices inconsistent with $\mathcal{T}$ can therefore not be both logically omniscient and also presuppose $\mathcal{T}$. The experimenter can, as shown by Proposition 5, identify which rules the subject understood, this is the largest sub-theory of $\mathcal{T}$ consistent with her behavior.

An agent who understands $\mathcal{T}$-implication—that is, whose preferences satisfy $\mathcal{T}$-**I**—believes the statements of $\mathcal{T}$ to be true: in any representation, $\lambda(t(\varphi)) = 1$ for $\varphi \in \mathcal{T}$.[7] Thus there is a relationship between an agent's knowledge and her understanding of implication. Under **I**, this relationship is one directional, as the following example points out: it is possible to believe a statement to be true, but not understand its implications.

Example 3. Let $\mathcal{L}$ be generated by two primitive propositions, $\textsc{p}$ and $\textsc{q}$. Let $\mathcal{S}_{\textsc{p}}$ be the closure of $\{\textsc{p}\}$ under implication and $\mathcal{S}_{\textsc{q}}$ the closure of $\{\textsc{q}\}$.

Now consider the model $\Omega = \{\omega\}$, $\lambda(\omega) = 1$, and t that maps $\mathcal{S}_{\textsc{p}}$ and $\mathcal{S}_{\textsc{q}}$ to Ω and all other statements to $\varnothing$. By construction, t is monotone. Also, by some interpretation the agent here modeled would *know* the theory $\{\textsc{p}\}$—she assigns probability 1 to (all direct entailments of) $\textsc{p}$. However, the agent would not be $\{\textsc{p}\}$-monotone, and therefore does not truly understand the theory.

To see this note that at ω, $\textsc{q}$ holds. Further, under the theory $\{\textsc{p}\}$ one can deduce from $\textsc{q}$ that $\textsc{p} \wedge \textsc{q}$ is true. However, $t(\textsc{p} \wedge \textsc{q}) = \varnothing \subsetneq t(\textsc{q})$. □

Strengthening the logical reasoning we require of the agent creates a tighter link between knowledge and understanding implication. In particular, if we require that the agent's preferences are represented by (Ω, t, λ) with λ additive and t not only monotone but $\wedge$-distributive, then the collection of

[7] This follows from the fact that $\mathbf{T} \overset{\mathcal{T}}{\Rightarrow} \varphi$, for $\varphi \in \mathcal{T}$, and the definition of $\mathcal{T}$-**I**.

statements the agent believes (i.e., assigns probability 1 to) is a theory—something that was not the case in Example 3 above. But more than that, as shown by the result below, this theory completely determines the set of implications the agent understands.

Proposition 6. *Let $\succcurlyeq$ satisfy* **NT**, **EU** *and* **IE**. *Let $\mathcal{T}$ be a theory and $\mathcal{S} \subseteq \mathcal{T}$ the unique largest sub-theory such that $\succcurlyeq$ satisfies $\mathcal{S}$-***I***. Then $\mathcal{S} = \{\varphi \in \mathcal{T} \mid b_\varphi \succcurlyeq b_{\mathbf{T}}\}$.*

While Proposition 5 ensures that under **I** there exists a largest sub-theory of the modeler's theory that coheres with the agents preference, it does nothing to explicitly construct it. If we a permit slightly strengthening the assumptions on the agent's reasoning capabilities, then this construction comes for free: Proposition 6 states that this sub-theory is exactly the intersection of the modeler's theory and the statements the agent believes to be true.

6 Discussion and Related Literature

Loosely speaking, the failures of contingent thinking discussed in the literature in psychology and economics can be classified into two types: failures of probabilistic reasoning and failures of logical reasoning. Explanations of the first type explain behavior by assuming that agents' subjective models of uncertainty do not conform to the usual rules of probability. Such models include ambiguity aversion (**??**), over weighting small probabilities (**??**), correlation neglect (**??**), etc.

Explanations of the second type posit that agents' subjective models of uncertainty do not conform to the usual rules of *logic*, specifically, they fail to perceive that some contingencies imply others. While experimental evidence has pointed towards explanations of this later type, general models of such behavior are less pervasive—perhaps because modeling an agent's understanding of implication seems more involved than generalizing probability measures.

?, in a series of experiments that serve as the basis for Example 1, discovered the *conjunction fallacy*, where subjects rank a statement of the form P ∧ Q as strictly more likely than the corresponding statement Q. Since this is a logical impossibility irrespective of the interpretation of propositions P and Q, their findings provide unambiguous evidence of systematic misperception of implications.

?, extending the idea behind the *winner's curse* (**?**), argue that subject's inclination to play dominated strategies arises from their failure to properly condition on the relevant contingencies; for example, a bidder in a common value auction failing to realize that conditional on submitting the highest bid, she also had the highest private signal. In these environments, subject's behavior can be explained by the thesis that they fail to perceive (or they ignore) the implications between various uncertainties (e.g., that winning implies a higher than average bid).

? examine such failure of contingent reasoning in the presence of uncertainty and "propose that aggregating over multiple possible values of the state is especially difficult when there is uncertainty." The connection between logical and probabilistic reasoning, as outlined in this paper, can provide a theoretical justification for this result: as errors in probabilistic thinking can often be equivalently characterized as errors in contingent thinking, it stands to reason that eliminating external probabilistic uncertainty will reduce the contingent errors made by subjects. Taken to the extreme, if an agent has a subjective model of uncertainty, (Ω, t, λ) with t exact, than certain statements (i.e., statements that are implied by $\mathbf{T}$) are always mapped to the entirety of Ω and assigned λ-probability 1. As such, in the absence of any uncertainty whatsoever, we should see no failure of contingent reasoning.[8]

? find that subject's deliberately randomize their actions across identi-

[8] Of course, even if all uncertainty is eliminated from the experimenter's perspective, there may be structural or background uncertainty harbored by the subject, see Footnote 6.

cal decision problems, and conclude “at least part of the mixing behavior observed in probability matching decisions comes from subjects’ difficulty in thinking contingently.” In line with findings above, they further see that subject’s randomize less when uncertainty about each decision problem is resolved sequentially, rather than only after all decisions have been made. This adds weight to the observation that higher exposure to uncertainty seems to exacerbate failures of contingent reasoning.

Our paper also contributes to the methodological literature on experimental design, as it provides a methodology for discerning between irrational subjects who have true violations of logical thinking from subjects who misunderstood or misinterpreted experimental instructions. **??** provide convincing evidence that observations often explained by behavioral biases might actually be the result of failing to understand instructions. Our results establish how to test for this in a more universal way, and show how to identify which instructions were misunderstood.

There are some models of strategic interaction, for example *cursed equilibrium*, (**?**) or *behavioral equilibrium* (**?**), that account for agents failing to understand implicative relation between other agent’s actions and their private information. For example cursed equilibrium “assumes that each player incorrectly believes that with positive probability each profile of types of the other players plays the same mixed action profile that corresponds to their average distribution of actions, rather than their true, type-specific action profile.”

? and **?** both explore the relation between contingent and probabilistic reasoning. The former paper shares much in common with ours, and shows how a decision maker’s understanding of implications can result in ambiguity aversion. In particular, the author considers an agent who entertains a refinement of the ‘objective’ state-space (i.e., the state space that governs payoffs), and maps individual states in her the objective state space into sets of states in her subjective refinement. **?** then shows that properties of this mapping can induce ambiguity aversion, in particular when the inverse mapping fails to be $\vee$-distributive.

? show that that failures of Savage's sure-thing-principle are inextricably related to failures of contingent reasoning (in the sense of being represented by contingent preferences). In addition they find experimental evidence of this relation, and write

> ...subjects are not good at thinking through the state space in the way modelers often assume and ... incomplete preferences or anomalies may precisely stem from the fact that states are not naturally given, may be hard to construct, or some states may not be salient.

We take this as strong motivation for the construction of a state-space as a representation object rather than as an objective and exogenous primitive.

Our paper is also relevant to the decision theoretic literature that dispenses with logical or probabilistic omniscience. While the probabilistic side has been well explored—see **?** for an overview—decision theoretic models of relaxed logical reasoning are less prevalent. **?** presents a model in which agents may consider impossible states of affairs. This violates our exactness axiom, although is permitted in our most general framework. Recently, in the economics literature, **?** explores belief updating in the face of contradictory claims. This follows on the extensive literature on *belief revision* centered in epistemic logical via philosophy and computer science: for examples see **?????**. More broadly, our paper shares much, both mathematically and philosophically, with logical representations of knowledge and awareness, see **?** for an overview.

Closely related to the methodology of our paper are models of syntactic decision theory, where the primitive relates to statements about the world rather than semantic acts or lotteries, principally **??** and **?** and to a lesser extent **??** and **?**. **?** deal with failures of logical implication, but, in contrast to the present paper, take the state space as a primitive object and are concerned with a particular (probabilistically sophisticated) representation of conditional judgements (i.e., the probability of φ given ψ). Like the model we present here, a key aspect of **?** is that the state-space is part of the representation rather than the primitive, so that the *interpretation* of uncertainty becomes purely subjective.

References

A Additional Representation Results

In this section we provide two additional axioms and corresponding restrictions on the representations. The later less interesting case, included simply for completeness, is that of full rationality, where t is sound and λ is additive. The more interesting and subtle case concerns a decision maker whose beliefs can be represented by a totally monotone capacity, often called a *belief function* following **?** and **?**.

Axiom 5—Inclusion/Exclusion (**IE**). Set $\{\varphi_1 \dots \varphi_n\}$ and let ψ be such that $\varphi_i \Rightarrow \psi$ for all $i \leq n$. Set Let E and O denote the (non-empty) subsets of $\{1, \dots n\}$ with even and odd numbers of elements, respectively. Moreover, for any $I \subseteq \{1, \dots n\}$ let φ_I be shorthand for $\bigwedge_{i \in I} \varphi_i$. Finally, set $m = \max\{|E|+1, |O|\}$. Then

$$\Big\{\frac{1}{m}b_\psi, \frac{1}{m}b_{\varphi_I}, (1 - \frac{|E|+1}{m})b_{\mathbf{F}}\Big\}_{I \in E} \succcurlyeq \Big\{\frac{1}{m}b_{\varphi_I}, (1 - \frac{|O|}{m})b_{\mathbf{F}}\Big\}_{I \in O}.$$

Under Axioms **NT** and **EU**, Axiom **IE** implies **I** (and hence **E**). This follows immediately from taking a singleton φ. Axiom **IE**, an admittedly somewhat mechanical reproduction of total monotonicity to our decision theoretic primitive, ensures that $\succcurlyeq$ might be represented by a sound t and totally monotone λ. The additional equivalence postulated below, however, is not trivial: representation by a totally monotone λ (and sound t) is equivalent to representation by a $\wedge$-distributive T and additive λ. In particular, this shows that if an agents preferences can be represented by a belief function, then in any additive representation the agent never considers contradictory statements possible simultaneously—and immediate consequence of $\wedge$-distributivity.

Proposition 7. *The following are equivalent:*

1. $\succcurlyeq$ *satisfies Axioms* **NT**, **EU**, *and* **IE**,

2. $\succcurlyeq$ *is represented by a triple* (Ω, t, λ), *with* λ *additive and* t *exact and* $\wedge$-*distributive, and,*

3. $\succcurlyeq$ *is represented by a triple* (Ω', t', λ'), *with* λ' *totally monotone and* t' *sound.*

Proof of Proposition 7. We will show (1) implies (3) implies (2) implies (1). Assume (1). Then by Since under Axioms **NT** and **EU** Axiom **IE** implies **I**, Proposition 3 requires that there exist a representation of $\succcurlyeq$, (Ω, t, λ), with λ monotone and t sound. Since t is sound, hence a Boolean homomorphism, $t(\mathcal{L})$ is a field of sets. Assume without loss of generality $\Sigma = t(\mathcal{L})$. We will show that in fact, λ is totally monotone.

Indeed, let $A_1 \dots A_n \in \Sigma$. Since $\Sigma = t(\mathcal{L})$, there exists φ_i such that $t(\varphi_i) = A_i$ for $i \leq n$. Set $\psi = \bigvee_{i \leq n} \varphi_i$. Clearly, $\varphi_i \Rightarrow \psi$, so Axiom **IE** therefore dictates that

$$\Big\{\frac{1}{m} b_{\psi}, \frac{1}{m} b_{\varphi_I}, (1 - \frac{|E| + 1}{m}) b_{\mathbf{F}}\Big\}_{I \in E} \succcurlyeq \Big\{\frac{1}{m} b_{\varphi_I}, (1 - \frac{|O|}{m}) b_{\mathbf{F}}\Big\}_{I \in O},$$

where E, I and m are defined in Axiom **IE**. Given the representation, where $\varphi_I = \bigwedge_{i \in I} \varphi_i$ we have

$$\frac{1}{m} \lambda(t(\psi)) + \sum_{I \in E} \frac{1}{m} \lambda(t(\varphi_I)) \geq \sum_{I \in O} \frac{1}{m} \lambda(t(\varphi_I))$$

which, given the logical omniscience of t provides $t(\psi) = \bigcup_{i \leq n} t(\varphi_i) = \bigcup_{i \leq n} A_i$ and $t(\varphi_I) = \bigcap_{i \in I} t(\varphi_i) = \bigcap_{i \in I} A_i$ is exactly the dictate of total monotonicity.

Now assume (3), that $\succcurlyeq$ is represented by (Ω', t', λ') with λ' totally monotone. Let Ω be the set of all non-empty subsets of Ω' and let Σ denote the algebra generated by the principal ideals of Ω': $\{\{A \in \Omega' | A \subseteq B\} \mid B \in \Omega'\}$.

By Theorem A of **?** there exists an additive $\lambda \in \Lambda(\Omega, \Sigma)$ such that

$$\lambda' = \int_{\Omega} u_A d\lambda(A)$$

where u_A is the capacity on Ω defined by

$$u_A(B) = \begin{cases} 1 \text{ if } A \subseteq B \\ 0 \text{ otherwise.} \end{cases}$$

Finally set $t(\varphi) = \{A \in \Omega \mid A \subseteq t'(\varphi)\}$. Owing to the logical omniscience of t', we have $t(\varphi \wedge \psi) = \{A \in \Omega \mid A \subseteq t'(\varphi) \cap t'(\psi)\} = t(\varphi) \cap t(\psi)$, so that t is

$\wedge$-distributive (and it is obliviously exact). Moreover,

$$\lambda(t(\varphi)) = \int_{\Omega} \mathbb{1}_{\{A\in\Omega|A\subseteq t'(\varphi)\}} d\lambda = \int_{\Omega} u_A(t'(\varphi)) d\lambda(A) = \lambda'(t'(\varphi)),$$

so (Ω, t, λ) represents $\succcurlyeq$.

Finally, assume (2), that $\succcurlyeq$ is represented by (Ω, t, λ) with λ additive and t exact and $\wedge$-distributive. By Remark 4, t is monotone. Take some $\{\varphi_1 \dots \varphi_n\}$ and let ψ be such that $\varphi_i \Rightarrow \psi$ for all $i \leq n$.

By the monotonicity of t, $t(\psi) \supseteq t(\varphi_i)$ for each i and hence $t(\psi) \supseteq \bigcup_{i\leq n} t(\varphi_i)$, indicating $\lambda(t(\psi)) \geq \lambda(\bigcup_{i\leq n} t(\varphi_i))$. Since λ is additive, we know that

$$\lambda(\bigcup_{i\leq n} t(\varphi_i)) = \sum_{\{I|\varnothing\neq I\subseteq\{1\dots n\}\}} (-1)^{|I|+1} \lambda\Big(\bigcap_{i\in I} t(\varphi_i)\Big) = \sum_{\{I|\varnothing\neq I\subseteq\{1\dots n\}\}} (-1)^{|I|+1} \lambda\Big(t(\varphi_I)\Big),$$

where $\varphi_I = \bigwedge_{i\in I} \varphi_i$, and the final equality comes form $\wedge$-distributivity. Setting $m = \max\{|E|+1, |O|\}$ we see that

$$\frac{1}{m}\lambda(t(\psi)) + \sum_{I\in E} \frac{1}{m}\lambda(t(\varphi_I)) \geq \sum_{I\in O} \frac{1}{m}\lambda(t(\varphi_I)),$$

where E and O are as defined in Axiom **IE**, which shows via the representation that Axiom **IE** holds. ■

And now, the axiom that delivers full rationality, reminiscent of so many independence type axioms:

Axiom 6—Additivity (**A**). If $\varphi \wedge \varphi' \Rightarrow \mathbf{F}$ then $\{\frac{1}{2}b_{\varphi}, \frac{1}{2}b_{\varphi'}\} \sim \{\frac{1}{2}b_{\varphi\vee\varphi'}, \frac{1}{2}b_{\mathbf{F}}\}$.

Under Axioms **NT** and **EU**, Axiom **A** implies **IE** (and hence **I** and **E**). However, this is a bit of a pain to prove directly—hint, its inductive—but it is obviously implied by our representation theorems. It is straightforward, to show that Axiom **A** implies **I**, and is the content of Lemma 1 in the proof of the main proposition.

Proposition 8. *The following are equivalent:*

1. *$\succcurlyeq$ satisfies Axioms* **NT**, **EU**, *and* **A**, *and,*
2. *$\succcurlyeq$ is represented by a triple (Ω, t, λ), with λ additive and t sound.*

Proof of Proposition 8. First, we prove the following simple lemma:

Lemma 1. *Under Axioms* **NT** *and* **EU***, Axiom* **A** *implies* **I***.*

Proof. Let $\varphi \Rightarrow \psi$. Set $\varphi' = (\psi \wedge \neg\varphi)$. Then we have $\varphi \vee (\psi \wedge \neg\varphi) \Leftrightarrow (\varphi \vee \psi) \wedge (\varphi \vee \neg\varphi) \Leftrightarrow \psi \wedge \mathbf{T} \Leftrightarrow \varphi$ and $\varphi \wedge (\psi \wedge \neg\varphi) \Leftrightarrow \mathbf{F}$. So, we have that $\{\frac{1}{2}b_\varphi, \frac{1}{2}b_{(\psi\wedge\neg\varphi)}\} \sim \{\frac{1}{2}b_\psi, \frac{1}{2}b_{\mathbf{F}}\}$, which given Axioms **NT** and **EU** clearly implies $b_\psi \succcurlyeq b_\varphi$. ★

By by the above lemma and Proposition 3, there exists a representation of $\succcurlyeq$, (Ω, t, λ), with λ monotone and t sound. Since t is sound, hence a Boolean homomorphism, $t(\mathcal{L})$ is a field of sets. Assume without loss of generality $\Sigma = t(\mathcal{L})$. We will show that in fact, λ is additive.

Indeed, let $A, B \in \Sigma$ with $A \cap B = \varnothing$. Since $\Sigma = t(\mathcal{L})$, there exists φ^A, φ^B such that $t(\varphi^A) = A$ and $t(\varphi^B) = B$. Set $\psi = \varphi^A \vee \varphi^B$. The soundness of t provides $t(\varphi^A \wedge \varphi^B) = A \cap B = \varnothing = t(\mathbf{F})$, so $\varphi^A \wedge \varphi^B \Leftrightarrow \mathbf{F}$. Axiom **A** therefore dictates that $\{\frac{1}{2}b_{\varphi^A}, \frac{1}{2}b_{\varphi^B}\} \sim \{\frac{1}{2}b_\psi, \frac{1}{2}b_{\mathbf{F}}\}$ or, given the representation

$$\frac{1}{2}\lambda(t(\varphi^A)) + \frac{1}{2}\lambda(t(\varphi^B)) = \frac{1}{2}\lambda(t(\psi)) + \frac{1}{2}\lambda(t(\mathbf{F}))$$

which, given that $t(\psi) = t(\varphi^A \vee \varphi^B) = t(\varphi^A) \cup t(\varphi^B)$ and $t(\mathbf{F}) = \varnothing$ is

$$\lambda(A) + \lambda(B) = \lambda(A \cup B),$$

showing that λ is additive. ■

B An application to Rationalizability in Games

B.1 Integral Representations

In this section we investigate the model under only our most minimal restrictions for a logically sound representation, that is, under Axioms **NT**, **EU**, and **E**. Even under the weak assumption **E**, we can unambiguously assign utility values to more general class of acts which map statements into outcomes.

Towards this, define a *strategy* to be mapping $s : \mathcal{L} \to \mathbb{R}$ with finite support. Notice that a primitive bet, defined in Section 3.2, is a special type of strategy, where the support is a singleton. Let S denote the set of strategies, and notice that S is a mixture space under pointwise mixtures.

Let MS denote the set of lotteries over S, referred to as mixed strategies. The set of all bets, $\mathcal{B}$, is contained inside of MS.

Let $\dot{\succcurlyeq}$ be a preference relation over MS, and let $\succcurlyeq$ denote its restriction to $\mathcal{B}$. Assume that (Ω, t, λ) represents $\succcurlyeq$, with t sound—thus, $\succcurlyeq$ satisfies **NT**, **EU**, and **E**. Under t, a strategy is naturally associated to a $t(\mathcal{L})$-measurable function from Ω to $\mathbb{R}$, via:

$$t_\circ : s \mapsto \sum_{\mathrm{supp}(s)} s(\varphi)\mathbb{1}_{t(\varphi)}. \tag{2}$$

Notice t° is the unique linear function satisfying $t_\circ(b_\varphi) = \mathbb{1}_{t(\varphi)}$. Say $\dot{\succcurlyeq}$ has an integral-representation if: $\dot{\succcurlyeq}$ is linear in mixtures and for all pure strategies, $s, s' \in S$, $s \dot{\succcurlyeq} s'$ if and only if

$$\int t_\circ(s)\mathrm{d}\lambda \geq \int t_\circ(s')\mathrm{d}\lambda.$$

Remark 4. Naturally, we might ask, when does $\dot{\succcurlyeq}$ have an integral representation? We briefly remark on the answer to this question, but for the sake of brevity, do not go into details. Under the map $t_\circ$ we can identify strategies as functions $\Omega \to \mathbb{R}$. Thus, we can place axioms directly on acts over Ω, and use $t_\circ^{-1}$ to pull back the restrictions to our primitive. On this space we can employ the usual Anscombe-Aumann type axiomatic structure.

Say that $\dot{\succcurlyeq}$ is *obvious dominant* if for all $x, x' : \Omega \to \mathbb{R}$ such that $\min_\Omega x(\omega) \geq \max_\Omega y'(\omega)$ we have $s \dot{\succcurlyeq} s'$ for all $s \in t_\circ^{-1}(x)$ and $s' \in t_\circ^{-1}(x')$. Then, in a straightforward generalization of **?**: $\dot{\succcurlyeq}$ is a continuous weak order, satisfying co-monotone independence and obvious dominance, if and only if it has an integral representation with a general λ (strengthening obvious dominance to monotonicity provides that λ is a capacity and further strengthening co-monotone independence to full independence provides λ is additive).

It is perhaps worth pointing out that our notion of obvious dominance coincides with that of **?**. □

To relate to the above established duality in the representation, we would like to extend our notion of an integral representation to (Ω', t', λ'), another representation of $\succcurlyeq$, with t' exact but not necessarily sound, and λ' additive. However, the mapping $t_\circ$ makes sense only for sound t. To see this consider the strategy s given by its support: $\varphi \mapsto 1$ and $\neg\varphi \mapsto 1$. Read in natural language, the strategy pays 1 if either φ or $\neg\varphi$, and 0 elsewhere, which should reasonably identified with the strategy s' given by $\varphi \vee \neg\varphi \mapsto 1$. However, if t' is not symmetric, for example if $t'(\varphi) \subsetneq \Omega \setminus t(\neg\varphi)$ then blindly applying (2) yields for s a function that is 1 on $t'(\varphi) \cup t(\neg\varphi) \neq \Omega$ and 0 elsewhere and for s' a constant function 1.

While it is certainly possible that the framing of a strategy could impact how a decision maker interprets it, in the paper we eschew such additional constraints on the agent's cognition. That is to say, we assume that when agent observers the strategy s paying the same across both φ and $\neg\varphi$ she identifies s with the strategy whose payment is contingent on $\varphi \vee \neg\varphi$. This assumption has the added benefit of being able to harness our earlier duality results for practical uses, as it allows the modeler freedom in her choice of representation.

Formally, let (Ω', t', λ'), with t exact and λ' additive, be a representation of $\succcurlyeq$. We will define a map $t'_\bullet : \mathbb{R}^{\mathcal{L}} \to \mathbb{R}^{\Omega'} / \overset{\lambda'}{\sim}$ where the co-domain is the equivalence classes of λ'-almost-everywhere equal functions.

Towards this, for any (Ω, t, λ), with t sound, consider any $t(\mathcal{L})$-measurable $x : \Omega \to \mathbb{R}$ map with finite image $\{\alpha_1, \dots, \alpha_n\}$, where $\alpha_k > \alpha_{k+1}$. For every $k = 1, \dots, n-1$ choose some $\varphi_k \in t^{-1} \circ x^{-1}(\{\alpha_1, \dots, \alpha_k\})$. Then, (defining $\alpha_{n+1} = 0$) set the map:

$$\xi^t_{t'} : x \mapsto \left[\sum_{k=1}^{n} (\alpha_k - \alpha_{k+1}) \mathbb{1}_{t'(\varphi_k)} \right], \tag{3}$$

where the brackets $[\cdot]$ indicate the $\overset{\lambda'}{\sim}$ equivalence class. This is well defined since if $\varphi, \psi \in t^{-1}(A)$, then $\lambda'(t'(\varphi)) = \lambda'(t'(\psi)) = \lambda(t(A))$ by definition of a representation so the resulting functions are λ'-almost-everywhere equal for any choice of $\varphi_1 \dots \varphi_k$.

Lemma 2. *The map $\xi^t_{t'} \circ t_\circ$ is unique (in t) λ'-almost everywhere.*

As a result of Lemma 2 we can define the map $t'_\bullet : \mathbb{R}^{\mathcal{L}} \to \mathbb{R}^{\Omega'} / \overset{\lambda'}{\sim}$ as $\xi^t_{t'} \circ t_\circ$ for any choice of representation (Ω, t, λ) with t sound.

Example 4. Let $\mathcal{L}$ be defined over the propositions p, q. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Let λ be the capacity defined by $\lambda(\{\omega_1\}) = \lambda(\{\omega_3\}) = \lambda(\{\omega_1, \omega_2\}) = \lambda(\{\omega_2, \omega_3\}) = \frac{1}{3}$ and $\lambda(\{\omega_2\}) = 0$ and $\lambda(\{\omega_1, \omega_3\}) = \frac{2}{3}$. Let t be the (sound) truth valuation defined by $t(p) = \{\omega_1, \omega_2\}$ and $t(q) = \{\omega_1\}$, and $\tilde{t}$ be the (sound) truth valuation defined by $\tilde{t}(p) = \{\omega_1\}$ and $\tilde{t}(q) = \{\omega_1, \omega_2\}$. Notice that both (Ω, t, λ) and $(\Omega, \tilde{t}, \lambda)$ represent the same $\succcurlyeq$.

$$\begin{tikzcd}
\mathbb{R}^{\mathcal{L}} \arrow[r, "t_\circ"] \arrow[d, "\tilde{t}_\circ"'] \arrow[dr, dashed, "t'_\bullet"] & \mathbb{R}^{\Omega} \arrow[d, "\xi^{t}_{t'}"] \\
\mathbb{R}^{\tilde{\Omega}} \arrow[r, "\xi^{\tilde{t}}_{t'}"] & \mathbb{R}^{\Omega'}/\overset{\lambda'}{\sim}
\end{tikzcd}$$

Figure 2: The map $t_\bullet$ is defined by the commutative square above.

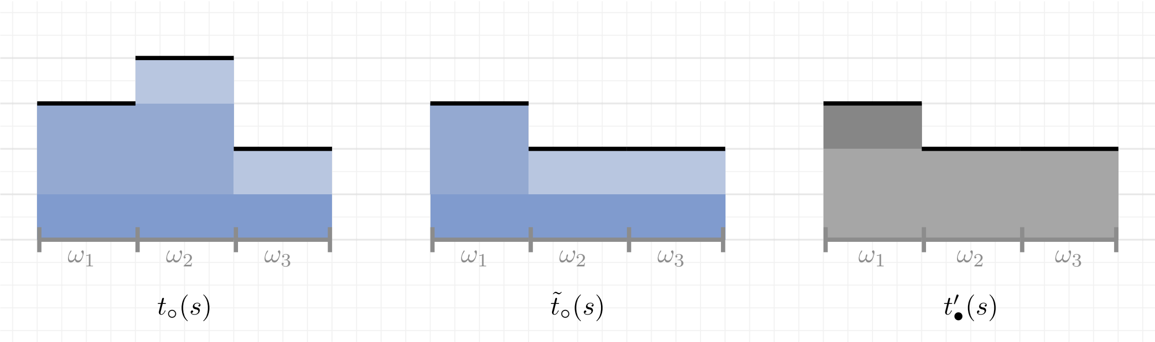

Figure 3: The maps $t_\circ(s)$, $\tilde{t}_\circ(s)$ and $t'_\bullet(s)$ from Example 4.

Further consider the exact, monotone truth valuation t' defined by $t'(p \wedge q) = t'(p \vee q) = t'(p) = t'(q) = \{w_1\}$ and $t'(\neg p \wedge \neg q) = t'(\neg p \vee \neg q) = t'(\neg p) = t'(\neg q) = \{w_2\}$. Then if λ' is the uniform (additive) measure on Ω, the (Ω, t', λ') is also a representation of $\succcurlyeq$.

Now consider the strategy $s : \mathbf{T} \mapsto 1; p \mapsto 2; \neg q \mapsto 1$. The maps $t_\circ(s)$ and $\tilde{t}_\circ(s)$ are given by the top half of Figure 3. Notice that

$$\begin{aligned}
\xi^{t}_{t'} \circ t_\circ(s) &= 2\mathbb{1}_{t'(T)} + \mathbb{1}_{t'(p)} + \mathbb{1}_{t'(p \wedge \neg q)} \\
&= 2\mathbb{1}_{\Omega} + \mathbb{1}_{\omega_1} + \mathbb{1}_{\varnothing} \\
&= 2\mathbb{1}_{t'(T)} + \mathbb{1}_{t'(p)} = \xi^{\tilde{t}}_{t'} \circ \tilde{t}_\circ(s),
\end{aligned}$$

so that our initial choice of model was irrelevant. □

Proposition 9. *Let (Ω, t, λ) and (Ω', t', λ') be a representations of $\succcurlyeq$ with*

t sound, t' exact and λ' additive. Then for all strategies $s : \mathcal{L} \to \mathbb{R}$,

$$\int t_\circ(s)\mathrm{d}\lambda = \int t'_\bullet(s)\mathrm{d}\lambda'.$$

Corollary 1. *Let $\dot{\succcurlyeq}$ admit an integral representation. Then (Ω, t, λ), with either t sound or t exact and λ additive, is an integral representation if and only if it is a representation of $\succcurlyeq$.*

B.2 Rationalizability

In this section, we use the above established results to examine the problem of determining rationalizability under non-expected utility beliefs. We assume a fixed $\mathcal{L}$ and the existence of a 'true' or 'objective' state-space Ω along with a sound truth valuation $t : \mathcal{L} \to \Omega$. With only slight hyperbole, this is the implicit set up of all economic applications of decision making under uncertainty. A strategy $s : \mathcal{L} \to \mathbb{R}$, in this setup, is often associated directly with its representation as a function over $\Omega \to \mathbb{R}$ via the (often implicit) map $t_\circ$.

Let $A \subseteq S$ denote a finite set of (pure) strategies that are available to the decision maker. Suppose that the decision maker can randomize over this set of strategies. Then the set of mixed strategies available is the convex hull of A, denoted by $co(A)$, a subset of MS.

Often a modeler will ask, after observing the choice $s \in A$ of the agent, was this choice rationalizable? In other words, is there a subjective model of uncertainty that the decision maker might hold that would yield s as the utility maximizing choice from A. Of course, since we have fixed both the state space and the truth valuation, the modeler's question reduces to the search for a $\lambda \in \Lambda(\Omega, t(\mathcal{L}))$ such that

$$\int t_\circ(s)\mathrm{d}\lambda \geq \int t_\circ(s')\mathrm{d}\lambda$$

for all $s' \in A$.

The celebrated result of **?** and **?** provide an elegant solution when we restrict ourselves to additive likelihood assessments.

Lemma 3 (Wald-Pearce). *Let $A \subset \mathbb{R}^\Omega$ be finite. Then $x \in A$ is a rationalizable via an additive measure if and only if x is pointwise undominated in $co(A)$ (where mixtures are taking pointwise).*

Of course, we may want to allow a larger class of likelihood estimates to serve as potential rationales which in turn might permit the rationalization of more strategies. The following simple example illustrates.

Example 5. Let $\mathcal{L}$ be defined over the single proposition p and set $\Omega = \{\omega_1, \omega_2\}$, and t the sound truth valuation given by $t(p) = \{\omega_1\}$. Consider three strategies: $s_1 = p \mapsto 1$, $s_2 = \neg p \mapsto 1$ and $s_3 : \mathbf{T} \mapsto \frac{1}{3}$. Using the above theorem we see there is no additive λ that rationalizes s_3, seen by the fact that $\frac{1}{2}t_\circ(s_1) + \frac{1}{2}t_\circ(s2) = (\frac{1}{2}, \frac{1}{2})$ strictly pointwise dominates $t_\circ(s_3) = (\frac{1}{3}, \frac{1}{3})$.

However, s_3 can be rationalized by relaxing the additivity requirement on the likelihood assessment, for example consider the monotone λ give by $\lambda(\omega_1) = \lambda(\omega_2) = \frac{1}{4}$. We see that

$$\int t_\circ(s_3)\mathrm{d}\lambda = \frac{1}{3} > \frac{1}{4} = \int t_\circ(s_1)\mathrm{d}\lambda = \int t_\circ(s_2)\mathrm{d}\lambda,$$

So s_3 is a best response to the subjective belief embodied by λ. □

Mathematically, when we allow for non-additive rationalizations, the convex hull of A is not the right space to search for dominance as we can find dominated strategies that are in fact best responses. However, our duality results above point out that if (Ω, t, λ) represents the agent's preferences then we can find an alternative (Ω', t', λ') with t exact and λ' additive. Over this space, the Wald-Pearce Lemma applies.

Along this line of thought, the modeler could take the 'objective' model (Ω, t) and transform it into a model (Ω', t') in which t need not be sound. Then, supposing an additive measure λ' on Ω' could be pulled back into an arbitrary λ on Ω such that (Ω, t, λ) and (Ω', t', λ') represent the same $\succcurlyeq$, the modeler could check for rationalization in the additive model. Of course, by Proposition 9, the strategies in question must be transformed via $t'_\bullet$.

The lingering concerns are then (i) the possible choices for (Ω', t') are so immense they are not even a set, and (ii) given a particular choice, can we transform λ' into λ in a suitably regular way so as to preserve representation. To circumvent these issues we restrict ourselves to the *maximal* model relative to the objective model (Ω, t): take as given the sound model (Ω, t) and for each $A \in t(\mathcal{L}) - \{\Omega, \varnothing\}$ let $\Omega^m_A = \{0_A, 1_A\}$ and $\Omega^m = \prod_{t(\mathcal{L})} \Omega^m_A$. Set $t^m : \varphi \mapsto 1_t(\varphi)$ (and $t(\mathbf{T}) \ni \varphi \mapsto \Omega, t(\mathbf{F}) \ni \varphi \mapsto \varnothing$). We call (Ω^m, t^m) the maximal model relative to (Ω, t). The construction in Proposition 2 shows that that every exact preference $\succcurlyeq$ is represented by a maximal model with an additive λ (in particular maximal relative to the canonical model).

Proposition 10. *Fix $\mathcal{L}$ and (Ω, t) with t sound. Let $A \subset S$ be finite. Then $s \in A$ is a rationalizable via a likelihood function if and only if $t^m_\bullet(s)$ is pointwise undominated in $co(t^m_\bullet(A))$.*

Example 5 (continued). Recall the model from earlier in the example: $\Omega = \{\omega_1, \omega_2\}$, and $t(p) = \{\omega_1\}$. Put the notation $A = t(p) = \omega_1$ and $A = t(\neg p) = \omega_2$. The maximal model for our $\Omega^m = \{(1_{A,B}), (1_A, 0_B), (0_A, 1_B), (0_A, 0_B)\}$ with $t : p \mapsto \{(1_{A,B}), (1_A, 0_B)\}$ and $\neg p \mapsto (1_A, 1_B), (0_A, 1_B)\}$.

The three strategies considered were $s_1 = p \mapsto 1$, $s_2 = \neg p \mapsto 1$ and $s_3 : \mathbf{T} \mapsto \frac{1}{3}$, which correspond to

$$t^m_\bullet(s_1) = \begin{cases} (1_A, 1_B) & \mapsto 1 \\ (1_A, 0_B) & \mapsto 1 \\ (0_A, 1_B) & \mapsto 0 \\ (0_A, 0_B) & \mapsto 0 \end{cases} \quad t^m_\bullet(s_2) = \begin{cases} (1_A, 1_B) & \mapsto 1 \\ (1_A, 0_B) & \mapsto 0 \\ (0_A, 1_B) & \mapsto 1 \\ (0_A, 0_B) & \mapsto 0 \end{cases} \quad t^m_\bullet(s_3) = \begin{cases} (1_A, 1_B) & \mapsto \frac{1}{3} \\ (1_A, 0_B) & \mapsto \frac{1}{3} \\ (0_A, 1_B) & \mapsto \frac{1}{3} \\ (0_A, 0_B) & \mapsto \frac{1}{3} \end{cases}$$

From this vantage, it is clear that s_3 is undominated as it is the only strategy that yields a positive payoff in state $(0_A, 0_B)$. Hence we can conclude immediately that there is a likelihood function that rationalizes s_3 in the original model. □

C Proofs omitted from the text

Lemma 4. *The following are true: (i) If t is exact and $\wedge$-distributive it is monotone. (ii) If t is exact, symmetric and $\wedge$-distributive it sound. (iii) If $h : \Omega \to \Omega'$ is a Boolean homomorphism, then $h \circ t : \mathcal{L} \to \Omega'$ is sound whenever t is.*

Proof of Lemma 4. (i) Let $\varphi \Rightarrow \psi$ then $\varphi \Leftrightarrow \varphi \wedge \psi$ so $t(\varphi) = t(\varphi \wedge \psi) = t(\varphi) \cap t(\psi)$ where the first equality follows from exactness and the second $\wedge$-distributivity. (ii) This follows from the fact that t is a homomorphism from $\mathcal{L}$, seen as the term algebra defined by the grammar, to the powerset of Ω, seen as a Boolean algebra, and that Boolean homomorphisms preserve logical validity. (iii) This is immediate from (ii). ■

Proof of Proposition 1. That (2) implies (1) is trivial. We first show the existence of a representation for primitive bets. By standard arguments, there exists, for each $\varphi \in \mathcal{L}$ a unique $\pi(\varphi) \in [0,1]$ such that $\pi(\varphi)b_{\mathbf{T}}+(1-\pi(\varphi))b_{\mathbf{F}} \sim b_\varphi$. For each $\varphi \in \mathcal{L} - \{\mathbf{T}, \mathbf{F}\}$ let $\Omega_\varphi = \{0_\varphi, 1_\varphi\}$ and let $\lambda_\varphi \in \Delta(\Omega_\varphi)$ be the measure such that $\lambda_\varphi(1_\varphi) = \pi(\varphi)$. Then a representation, for primitive bets, follows by setting $\Omega = \prod_{\mathcal{L}} \Omega_\varphi$, $t : \varphi \mapsto 1_\varphi$ (and $\mathbf{T} \mapsto \Omega, \mathbf{F} \mapsto \varnothing$) and λ to the corresponding product measure. Again, the standard inductive argument extends this representation to arbitrary bets using Independence. ■

Proof of Proposition 2. That (3) implies (1) straightforward. That (1) implies (2) follows from a slight variant of the proof of proposition 1, where $\mathcal{L}$ is quotiented by logical equivalence. We show that (2) implies (3).

Let (Ω, t, λ) be a representation given by (2), with λ additive. Set $[\varphi] = \{\psi \in \mathcal{L} \mid \varphi \Leftrightarrow \psi\}$. By the Stone representation theorem, $\mathbf{LT}(\mathcal{L})$ is embeddable in the powerset of a some set Ω', by some Boolean-homomorphism, h. Set $\Sigma = h(\mathbf{LT}(\mathcal{L}))$, which is a field of sets. Let $t' : \mathcal{L} \to \Omega'$ be given by $t' : \varphi \mapsto h([\varphi])$. By Lemma 4(iii), t' is sound. Set $\lambda'(t'(\varphi)) = \lambda(t(\varphi))$ which is well defined since t is exact. ■

Proof of Proposition 3. That (2) implies (1) straightforward. We first show that (1) implies (3). Since Axiom **I** trivially implies Axiom **E**, we have that $\succcurlyeq$ is represented by a triple (Ω', t', λ'), with t sound. Since t is sound, hence a Boolean homomorphism, $t(\mathcal{L})$ is a field of sets. Assume without loss of generality $\Sigma = t(\mathcal{L})$. We will show that in fact, λ' is monotone. Indeed, let $A, B \in \Sigma$ with $A \subseteq B$. Since $\Sigma = t(\mathcal{L})$, there exists φ^A, φ^B such that $t(\varphi^A) = A$ and $t(\varphi^B) = B$. Logical omniscience provides $t(\varphi^A \wedge \varphi^B) = A \cap B = A$. Moreover, it is a logical tautology that $(\varphi^A \wedge \varphi^B) \Rightarrow \varphi^B$: so by Axiom **I**, $b_{\varphi^B} \succcurlyeq b_{\varphi^A \wedge \varphi^B}$, and hence $\lambda'(t(\varphi^B)) \geq \lambda'(t(\varphi^A \wedge \varphi^B))$, or $\lambda'(B) \geq \lambda'(A)$.

Now we show that (3) implies (2). Let (Ω', t', λ') be a representation given by (3). Let $\Omega = [0,1]$, Σ the Borel σ-algebra, and $\lambda \in \Lambda([0,1], \Sigma)$ the Lebesgue measure on $[0,1]$. Let t be the map $t : \varphi \mapsto [0, \nu(t'(\varphi))]$ (whenever $\lambda'(t'(\varphi)) > 0$ and $\varnothing$ otherwise). Let $\varphi \Rightarrow \psi$. Then since t' is sound $t'(\varphi) \subseteq t'(\psi)$, and since λ' is monotone, $\lambda'(t'(\varphi)) \leq \lambda'(t'(\psi))$. Hence $t(\varphi) = [0, \nu(t'(\varphi))] \subseteq [0, \nu(t'(\psi))] = t(\psi)$. So t is monotone as desired. ■

Proof of Proposition 4. The only if direction is trivial. So assume $\varphi \Rightarrow \psi$ and $b_\psi \succcurlyeq b_\varphi$. By Proposition 1, there exists some (Ω', t', λ') representing $\succcurlyeq$ with λ additive. Let $\Omega = [0,1]$, Σ the Borel σ-algebra, and $\lambda \in \Lambda([0,1], \Sigma)$ the Lebesgue measure on $[0,1]$. Let t be the map $t : \varphi \mapsto [0, \lambda'(t'(\varphi))]$ (whenever $\lambda'(t'(\varphi)) > 0$ and $\varnothing$ otherwise). Since $b_\psi \succcurlyeq b_\varphi$ we have $\lambda'(t'(\psi)) \geq \lambda'(t'(\psi))$ hence $t(\varphi) \subseteq t(\psi)$, and so such a model exists. ■

Proof of Proposition 5. Let $\mathbf{S} = \{\mathcal{S}' \subseteq \mathcal{T} \mid \succcurlyeq \text{ satisfies } \mathcal{S}'\text{-}\mathbf{I}\}$. Let $\mathcal{S} = \bigcup_{\mathbf{S}} \mathcal{S}' \subseteq \mathcal{T}$. (Since $\succcurlyeq$ satisfies Axiom **I**, the set $\mathbf{S}$ contains the set of all logical tautologies, and so is non-empty.) The proposition follows if $\succcurlyeq$ satisfies $\mathcal{S}$-**I**.

Let $\varphi_0 \overset{\mathcal{S}}{\Rightarrow} \psi$. Definitionally, there is a finite collection of statements $\{\eta_1 \ldots \eta_n\} \subseteq \mathcal{S}$ such that $\varphi_0 \bigwedge_{1 \leq i \leq n} \eta_i \Rightarrow \psi$. For each $1 \leq i \leq n$, let $\mathcal{S}_i \in \mathbf{S}$ contain η_i and set $\varphi_i = \varphi \bigwedge_{j \leq i} \eta_j$. Then, for each $1 \leq i \leq n$, we have $\varphi_{i-1} \overset{\mathcal{S}_i}{\Rightarrow} \varphi_i$.

We have $b_\psi \succcurlyeq b_{\varphi_n} \succcurlyeq b_{\varphi_{n-1}} \ldots \succcurlyeq b_{\varphi_0}$. The first preferential relation comes from Axiom **I** and the others by applying $\mathcal{S}_i$-**I** for each i. By the transitivity of $\succcurlyeq$, $b_\psi \succcurlyeq b_{\varphi_0}$, so $\succcurlyeq$ satisfies $\mathcal{S}$-**I** as desired. ■

Proof of Proposition 6. That $\mathcal{S} \subseteq \{\varphi \in \mathcal{T} \mid b_\varphi \succcurlyeq b_{\mathbf{T}}\}$ follows immediately from $\mathcal{S}$-**I** the fact that $\mathbf{T} \overset{\mathcal{S}}{\Rightarrow} \varphi$ for all $\varphi \in \mathcal{S}$. We will here show the converse.

First, let (Ω, t, λ), with λ additive and t exact and $\wedge$-distributive, represent $\succcurlyeq$, the existence of which is guaranteed by Proposition 7. Let $\varphi \in \mathcal{T}$ be such that $b_\varphi \succcurlyeq b_{\mathbf{T}}$. Let $\mathcal{S}' = \mathcal{S} \cup \{\varphi\}$. The proposition follows if $\succcurlyeq$ satisfies $\mathcal{S}'$-**I**, since this would imply, by the definition of $\mathcal{S}$, that $\mathcal{S}' \subseteq \mathcal{S}$.

Let $\psi \overset{\mathcal{S}'}{\Rightarrow} \eta$. Definitionally, this means there is a (possibly tautological) statement $\zeta \in \mathcal{S}$ such that $\psi \wedge \varphi \wedge \zeta \Rightarrow \eta$. The converse already established that $b_\zeta \succcurlyeq b_{\mathbf{T}}$. So we have that $\lambda(t(\zeta)) = \lambda(t(\varphi)) = 1$. Since t is $\wedge$-distributive, we have that $t(\psi \wedge \varphi \wedge \zeta) = t(\psi) \cap t(\varphi) \cap t(\zeta)$. Then, the additivity of λ provides $\lambda(t(\psi \wedge \varphi \wedge \zeta)) = \lambda(t(\psi))$.

Moreover, since $\psi \wedge \varphi \wedge \zeta \Rightarrow \eta$ and t is exact and $\wedge$-distributive, hence monotone, we have $t(\psi) \supseteq t(\psi \wedge \varphi \wedge \zeta)$, indicating $\lambda(t(\eta)) \geq \lambda(t(\psi \wedge \varphi \wedge \zeta)) = \lambda(t(\psi))$ and thus that $b_\eta \succcurlyeq b_\psi$. So $\succcurlyeq$ satisfies $\mathcal{S}'$-**I** as desired. ■

Proof of Lemma 2. Let (Ω, t, λ) be a representation of $\succcurlyeq$ with t sound. Take $s : \mathcal{L} \to \mathbb{R}$ and set $x = t_\circ(s)$ with range $\{\alpha_1, \ldots, \alpha_n\}$ where $\alpha_k > \alpha_{k+1}$.

Now, consider the set $\mathbf{B} = \{\text{supp}(s) \cap \mathcal{T} \mid \mathcal{T} \text{ is a maximal theory of } \mathcal{L}\}$.[9] Enumerate $\mathbf{B}$ as $B_1 \ldots B_m$ and let $\varphi_i = \bigwedge_{B_i} \varphi$ and $\beta_k = \sum_{B_i} s(\varphi)$. Without loss of generality assume that $\beta_i \geq \beta_{i+1}$.

It is well known that if $t : \mathcal{L} \to \Omega$ is sound then for all $\omega \in \Omega$, $M_\omega = \{\varphi \in \mathcal{L} \mid \omega \in t(\varphi)\}$ is maximally consistent and hence $x(\omega) = \sum_{\text{supp}(s) \cap M_\omega} s(\varphi) = \beta_i$ for some $i \leq m$.

[9] Recall a theory $\mathcal{T}$ is *maximal* if it is a theory—it is closed under implication and $\mathbf{F} \notin \mathcal{T}$—and for every $\varphi \in \mathcal{L}$ either $\varphi \in \mathcal{T}$ or $\neg\varphi \in \mathcal{T}$. Alternatively, the maximal theories of $\mathcal{L}$ are exactly the ultrafilters on $\mathcal{L}$ under the partial order $\Rightarrow$.

For each $i \leq m$, set $\psi_i = \bigvee\{\varphi_j \mid \beta_j \geq \beta_i\}$. Now if $\beta_i = \alpha_k$ for some k, then by construction $\psi_i \in t^{-1} \circ x^{-1}(\{\alpha_1, \ldots, \alpha_k\})$ and so $\lambda'(t'(\psi_i)) = \lambda'(E_k)$. Further, if $\beta_i \notin \{\alpha_1, \ldots, \alpha_n\}$, then $t(\psi) = \varnothing$ so in particular, $\lambda'(t'(\psi_i)) = 0$. Thus the function

$$\sum_{i=1}^{m} (\beta_i - \beta_{i-1}) \mathbb{1}_{t'(\psi_i)},$$

which does not depend on t, is λ'-almost everywhere equal to $\xi^t_{t'} \circ t_\circ$. ■

Proof of Proposition 9. Follows from definitions, where $\{\alpha_1, \ldots, \alpha_n\}$, with $\alpha_k > \alpha_{k+1}$, is the range of $t_\circ(s)$:

$$\begin{aligned}\int t'_\bullet(s) \mathrm{d}\lambda' &= \int \xi^t_{t'}(t_\circ(s)) \mathrm{d}\lambda' \\ &= \sum_{k=1}^{n} (\alpha_k - \alpha_{k+1}) \lambda'(t'(\varphi_k)) \\ &= \sum_{k=1}^{n} (\alpha_k - \alpha_{k+1}) \lambda(t(\varphi_k)) \\ &= \int t_\circ(s) \mathrm{d}\lambda'\end{aligned}$$

Where only the third equality is not definitional, and follows from the fact that both (Ω, t, λ) and (Ω', t', λ') are representations of $\succcurlyeq$. ■

Proof of Proposition 10. First, assume that $s \in A$ is a rationalizable via (Ω, t, λ) for a likelihood function λ so that

$$\int t_\circ(s) \mathrm{d}\lambda \geq \int t_\circ(s') \mathrm{d}\lambda, \tag{4}$$

for all $s' \in A$. Let $\succcurlyeq$ be the preference represented by (Ω, t, λ);

Let $\lambda^m_A \in \Lambda(\Omega^m_A)$ be the measure $\lambda^m_A(1_A) = \lambda(A)$, and λ^m the corresponding product measure over Ω^m. By construction, $\lambda^m(t^m(\varphi)) = \lambda^m(1_{t(\varphi)}) = l(t(\varphi))$, so that $(\Omega^m, t^m, \lambda^m)$ represents $\succcurlyeq$. Proposition 9 indicates

$$\int t^m_\bullet(s) \mathrm{d}\lambda^m \geq \int t^m_\bullet(s') \mathrm{d}\lambda^m$$

for all $s' \in A$. So by the Wald-Pearce Lemma, $t^m_\bullet(s)$ is undominated in $co(t^m_\bullet(A))$.

Now, assume $t^m_\bullet(s)$ is pointwise undominated in $co(t^m_\bullet(A))$. By the Wald-Pearce Lemma, we have the existence of an additive λ^m such that $(\Omega^m, t^m, \lambda^m)$ rationalizes $t^m_\bullet(s)$: that is

$$\int t^m_\bullet(s)\mathrm{d}\lambda^m \geq \int t^m_\bullet(s')\mathrm{d}\lambda^m \tag{5}$$

for all $s' \in A$. Let $\succcurlyeq$ be the preference represented by $(\Omega^m, t^m, \lambda^m)$.

Construct $\lambda \in \Lambda(\Omega, t(\mathcal{L}))$ by $\lambda(A) = \lambda^m(1_A)$. For any $\varphi \in \mathcal{L}$, by construction, $\lambda(t(\varphi)) = \lambda^m(1_{t(\varphi)}) = \lambda^m(t^m(\varphi))$ so (Ω, t, λ) represents $\succcurlyeq$. Thus, by Proposition 9, applied to (5), we have

$$\int t_\circ(s)\mathrm{d}\lambda \geq \int t_\circ(s')\mathrm{d}\lambda,$$

for all $s' \in A$: $s \in A$ is a rationalizable. ■

Пьермонт, Э., Суасо-Гарин, П.

Ошибки условного мышления [Электронный ресурс] : препринт WP9/2021/06 / А. Пента, П. Суасо-Гарин ; Нац. исслед. ун-т «Высшая школа экономики». – Электрон. текст. дан. (11,3 Мб). – М. : Изд. дом Высшей школы экономики, 2021. – (Серия WP9 «Исследования по экономике и финансам»). – 43 с. (На англ. яз.)

Мы предлагаем теоретическую основу для анализа агента, который неверно интерпретирует или неверно воспринимает истинную проблему принятия решения, с которой он сталкивается. В рамках этой концепции мы показываем, что широкий спектр поведения, наблюдаемого в экспериментальных условиях, проявляется как неспособность воспринимать последствия, другими словами, как должным образом учитывать логические отношения между различными непредвиденными обстоятельствами, относящимися к выигрышу. Мы представляем поведенческие характеристики, соответствующие нескольким критериям логической сложности, и показываем, как можно определить, какие последствия агент не осознает. Таким образом, наша структура предоставляет как методологию оценки уровня условного мышления агента, так и стратегию определения ее убеждений при отсутствии полной рациональности.

Пьермонт Эван, Ройял Холлоуэй, департамент экономики, Великобритания; E-mail: evan.piermont@rhul.ac.uk

Суасо-Гарин Пейо, МИЭФ, Национальный исследовательский университет «Высшая школа экономики», Российская Федерация; E-mail: p.zuazogarin@hse.ru

Препринты Национального исследовательского университета «Высшая школа экономики» размещаются по адресу: http://www.hse.ru/org/hse/wp

Препринт WP9/2021/06
Серия WP9
Исследования по экономике и финансам

Пьермонт Эван, Суасо-Гарин Пейо

Ошибки условного мышления

(*на английском языке*)

Изд. № 2340