

Uncertainty in Mechanism Design*

Giuseppe Lopomo[†]
Fuqua School of Business
Duke University

Luca Rigotti[‡]
Fuqua School of Business
Duke University

Chris Shannon[§]
Department of Economics
UC Berkeley

October 2009

Abstract

We consider mechanism design problems with Knightian uncertainty formalized using incomplete preferences, as in Bewley (1986). Without completeness, decision making depends on a set of beliefs, and an action is preferred to another if and only if it has larger expected utility for all beliefs in this set. We consider two natural notions of incentive compatibility in this setting: maximal incentive compatibility requires that no strategy has larger expected utility than reporting truthfully for all beliefs, while optimal incentive compatibility requires that reporting truthfully has larger expected utility than all other strategies for all beliefs. In a model with a continuum of types, we show that optimal incentive compatibility is equivalent to ex-post incentive compatibility under fairly general conditions on beliefs. In a model with a discrete type space, we characterize full extraction of rents generated from private information. We show that full extraction is generically possible with maximal incentive compatible mechanisms, but requires sufficient disagreement across types, which neither holds nor fails generically, with optimal incentive compatible mechanisms.

JEL Codes: D0, D5, D8, G1

Keywords: Knightian uncertainty, mechanism design, auctions, incomplete preferences.

*We thank David Ahn, Dino Gerardi, Botond Koszegi, Paul Milgrom, Ben Polak, Debraj Ray, and Ennio Stacchetti for helpful discussions and comments.

[†]The Fuqua School of Business, Duke University; glopomo@duke.edu

[‡]The Fuqua School of Business, Duke University; rigotti@duke.edu

[§]Department of Economics, UC Berkeley; cshannon@econ.berkeley.edu

1 Introduction

In his classic work, Knight (1921) suggests a distinction between uncertainty and risk, arguing that while risky events have known probabilities, the likelihood of uncertain events is more qualitative in nature and cannot be computed precisely. Ellsberg (1961) advances a more precise definition of uncertainty, in which an event is uncertain or ambiguous if it has unknown probability. Choice behavior reflecting this difference is inconsistent with the standard expected utility model, an observation that has inspired a significant amount of recent research in economics and decision theory. The extent to which Knightian uncertainty affects economic institutions rests on whether the ambiguity individuals perceive about the environment translates into equilibrium effects. We develop a simple extension of the standard mechanism design framework, and use this framework to study if and how Knightian uncertainty influences economic outcomes in settings with private information.

Following Harsanyi (1967), we assume that an agent's type encodes beliefs regarding the environment in which she acts. We depart from the standard model by assuming these beliefs are described by a *set* of probability distributions to reflect the presence of Knightian uncertainty.¹ Our model is in the spirit of the decision theory developed by Bewley (1986), in which agents' preferences may be incomplete. Without completeness, a state-contingent consumption bundle is preferred to another if and only if it yields larger expected utility for all probability distributions in some set of probabilities. When Knightian uncertainty is ruled out, this set becomes a singleton and the model reduces to standard expected utility.

We consider a mechanism design setup with a single agent whose utility depends on her type, the realized state of the world, and the outcome of the mechanism; the mechanism is itself a function of the realized state and the report of the agent.² As usual, we focus on conditions such that the mechanism induces the agent to reveal her private information, represented by her type. Without Knightian uncertainty, interim incentive compatibility requires that reporting one's type truthfully yields expected utility at least as large as any other strategy. With Knightian uncertainty, the beliefs of each type are described by a set of probability distributions, and a strategy has no unique expected utility value associated with it. Thus, there are two natural and distinct notions of incentive compatibility. A weak notion requires that no strategy yields higher expected utility than truthful reporting for all of the agent's beliefs; we call this *maximal incentive compatibility*. In a maximal incentive compatible mechanism, reporting truthfully might not be comparable to misreporting. A stronger notion of interim incentive compatibility requires instead that truth-telling yields expected utility at least as large as any other strategy for all possible beliefs; we call this *optimal incentive compatibility*. In an optimal incentive compatible mechanism, reporting truthfully is preferred to misreporting. In this sense, optimal incentive compatible mechanisms are robust to the presence of Knightian uncertainty.³

¹Ahn (2007) develops a theory of interactive beliefs allowing for agents to hold a compact set of beliefs at any level, and constructs a corresponding universal type space.

²Since the state can include the types of other players, this framework can be extended to mechanisms with many players.

³Optimal incentive compatibility is formally equivalent to incentive compatibility in a setup in which there is no Knightian uncertainty, but the designer has more limited information than in the usual model: he only knows

Our main theorem provides conditions under which optimal incentive compatibility is equivalent to ex post incentive compatibility. These conditions include standard regularity assumptions often invoked in mechanism design, such as a continuum of types and smoothness of the utility function. They also include novel restrictions regarding the richness of the agent’s beliefs. In particular, these conditions are satisfied if the correspondence mapping types into beliefs is lower hemi-continuous and has non-empty relative interior. Loosely speaking, this controls the way in which beliefs change as types change and requires that there is Knightian uncertainty about every state of the world. In the standard framework, when only risk is allowed, interim and ex-post incentive compatible mechanisms are in general very different. Knightian uncertainty, instead, can result in sharp restrictions on feasible mechanisms, even when this uncertainty is arbitrarily small. It generates a significant discontinuity with respect to standard Bayesian mechanism design and therefore it influences equilibrium outcomes.

The second part of the paper focuses on information rents. While the seminal work of Akerlof (1970) shows that asymmetric information can have welfare consequences, recent papers suggest that appropriately constructed mechanisms can extract all rents agents derive from private information. Crémer and McLean (1985), Crémer and McLean (1988), and McAfee and Reny (1992) show that in the standard Harsanyi framework, correlation in beliefs across types allows the designer to extract all of the surplus in a wide array of settings. Since beliefs are generically correlated, these results suggest that private information typically has no value.⁴

We provide necessary and sufficient conditions for full extraction of rents when Knightian uncertainty is allowed. We focus on settings with a discrete type space since with a continuum of types our main theorem shows that optimal incentive compatibility is often equivalent to ex post incentive compatibility, and well-known results imply that full extraction is not possible in that case. In contrast, with a discrete type space, full extraction may still be possible. We show that with a maximal incentive compatible mechanism, full extraction follows from the familiar condition of correlation in beliefs across types, applied to *some selection* from the agent’s sets of beliefs. In contrast, full extraction with an optimal incentive compatible mechanism requires that this correlation condition holds *uniformly* across all beliefs. When uncertainty is sufficiently large, full extraction under optimal incentive constraints becomes impossible since belief sets must eventually intersect. As a consequence of these results, we show that with Knightian uncertainty full extraction is neither generically possible nor generically impossible.

Our paper is related to different strands of literature in mechanism design and game theory. First, our theorem on the equivalence between ex post and interim incentive compatibility under Knightian uncertainty is related to the recent body of work on “detail-free” mechanisms and robustness in mechanism design. Much of this work is motivated by relaxing aspects of standard mechanism design assumptions like common knowledge and restrictions on higher order beliefs. For example, Bergemann and Morris (2005) model robustness to higher order beliefs by requiring implementation in the universal type space, and show that this is equivalent to ex post implementation in many settings. This result can be viewed as a generalization of earlier work in Ledyard (1978) and Ledyard (1979) using the modern language of the universal type space. In these

that the agent’s beliefs belong to some set, but does not know which element of this set represents them.

⁴Neeman (2004) and Heifetz and Neeman (2006) argue that correlation can also be thought of as restrictive.

papers, valuations are fixed as beliefs vary from type to type. By taking the resulting union over all possible beliefs, this setup has the flavor of our model when mechanisms must satisfy optimal incentive compatibility for the special case in which each type’s beliefs display an extreme form of Knightian uncertainty: belief sets contain all probability measures. In contrast, our main theorem applies even for arbitrarily small belief sets, and shows that in many common mechanism design settings, robustness to subtle and arbitrarily small variation in beliefs nonetheless requires ex post incentive compatible mechanisms.

A second strand of related literature deals with mechanism design with Knightian uncertainty and complete preferences. Chung and Ely (2007) consider robustness of auction mechanisms with respect to the designer’s beliefs about the agents’ types. Focusing on optimal mechanisms as the designer’s beliefs vary, Chung and Ely (2007) show that there exists at least one belief for which ex-post incentive compatible mechanisms are optimal. Bose, Ozdenoren, and Pape (2006) study optimal auction design when the bidders and the seller may be ambiguity averse in the sense of Gilboa and Schmeidler (1989) and Schmeidler (1989).⁵ Their main result shows that for an ambiguity neutral seller, the optimal auction must provide full insurance to all types of all bidders provided the bidders’ belief sets contain the seller’s (unique) belief. In contrast, we do not model the designer’s beliefs explicitly, and focus instead on robustness to models of individual agents’ beliefs.

Our results on the existence of information rents and the potential for full extraction are related to recent work emphasizing robustness. Neeman (2004) points out that the possibility of full extraction hinges critically on the assumption that types with different values have different beliefs. Heifetz and Neeman (2006) argue that this assumption is not satisfied generically in appropriate type spaces allowing for richer higher order beliefs. Our approach instead considers robustness to the introduction of Knightian uncertainty in simple type spaces. Our results show that the presence of Knightian uncertainty provides an alternative justification for the impossibility of full rent extraction in private information settings.

The paper proceeds as follows. Section 2 describes the decision-theoretic framework that motivates our model. Section 3 introduces the setup. Section 4 develops the equivalence between optimal and ex-post incentive compatibility. In Section 5 we study the problem of full extraction of information rents in Knightian mechanisms. Section 6 concludes.

2 Preliminaries: Incomplete Preferences and Uncertainty

In this section we briefly illustrate the Knightian decision theory presented in Bewley (1986), as it provides the foundation for our analysis.⁶ The main result in Bewley (1986) shows that a strict preference relation that is not necessarily complete, but satisfies all other axioms of the standard Anscombe-Aumann framework, can be represented by a utility index and a set of probability distributions. Incompleteness is thus reflected in multiplicity of beliefs: the unique

⁵See also Chen, Katuscak, and Ozdenoren (2007) and Levin and Ozdenoren (2004) for other work on auctions with ambiguity averse bidders.

⁶Bewley’s original paper has been published recently as Bewley (2002).

subjective probability distribution of the standard expected utility framework is replaced by a set of probability distributions.

Let $x = (x_1, \dots, x_S)$ and $y = (y_1, \dots, y_S)$ denote state-contingent payoff vectors in \mathbf{R}_+^S . Bewley (1986) gives axioms under which a preference relation \succ that is not necessarily complete can be represented by a closed, convex set Π of probability distributions on S and a continuous, concave function $u : \mathbf{R}_+ \rightarrow \mathbf{R}$, unique up to positive affine transformations, such that

$$x \succ y \quad \text{if and only if} \quad \sum_{s=1}^S \pi_s u(x_s) > \sum_{s=1}^S \pi_s u(y_s) \text{ for all } \pi \in \Pi.$$

Abusing notation slightly, we rewrite this as

$$x \succ y \quad \text{if and only if} \quad E_\pi [u(x)] > E_\pi [u(y)] \text{ for all } \pi \in \Pi,$$

where $E_\pi[\cdot]$ denotes the expectation with respect to the probability distribution $\pi = (\pi_1, \dots, \pi_S)$.⁷ If \succ is complete, the set Π reduces to a singleton and the usual expected utility representation obtains.⁸ Without completeness, comparisons between alternatives are carried out one probability distribution at a time: one bundle is strictly preferred to another if and only if its expected utility is larger under every probability distribution in the set Π .⁹

Bewley (1986) suggests that the above representation captures the Knightian distinction between risk and uncertainty, where an event is risky if its probability is known, and uncertain otherwise. The decision maker perceives only risk when Π is a singleton, and uncertainty otherwise. Thus incompleteness and uncertainty are two sides of the same phenomenon; both the amount of uncertainty that the decision maker perceives and the degree of incompleteness of her preference order \succ are measured by the size of the set Π .¹⁰

A picture may help to clarify this representation. In Figure 1, the axes measure utility levels in each of two possible states. A single probability distribution from the set Π determines the slope of an indifference set through the fixed bundle x associated with the corresponding expected utility function for that distribution; since the axes measure utility levels, these indifference sets are straight lines. Selecting different elements of the set Π generates a family of indifference sets through x , each corresponding to a different probability distribution. The thick lines represent the extreme elements of this family, while the thin ones represent other elements.

⁷Incompleteness in decision making in a von Neumann Morgenstern setting was first studied by Aumann (1962) and Aumann (1964). Recently, this work has been extended and clarified by Dubra, Maccheroni, and Ok (2004), Ok (2002) and Shapley and Baucells (2008). Preferences of this kind have also been studied by Ghirardato, Maccheroni, Marinacci, and Siniscalchi (2003), Gilboa, Maccheroni, Marinacci, and Schmeidler (2008), and axiomatized by Girotto and Holzer (2005) in an infinite state space.

⁸We say \succ is *complete* if for all $x \in \mathbf{R}_+^S$, $\text{cl} \{y \in \mathbf{R}_+^S : x \succ y \text{ or } y \succ x\} = \mathbf{R}_+^S$.

⁹The natural notion of indifference defines two bundles to be indifferent if they have the same expected utility for each probability distribution in Π .

¹⁰For precise results along these lines, see Ghirardato, Maccheroni, and Marinacci (2004) or Rigotti and Shannon (2005).

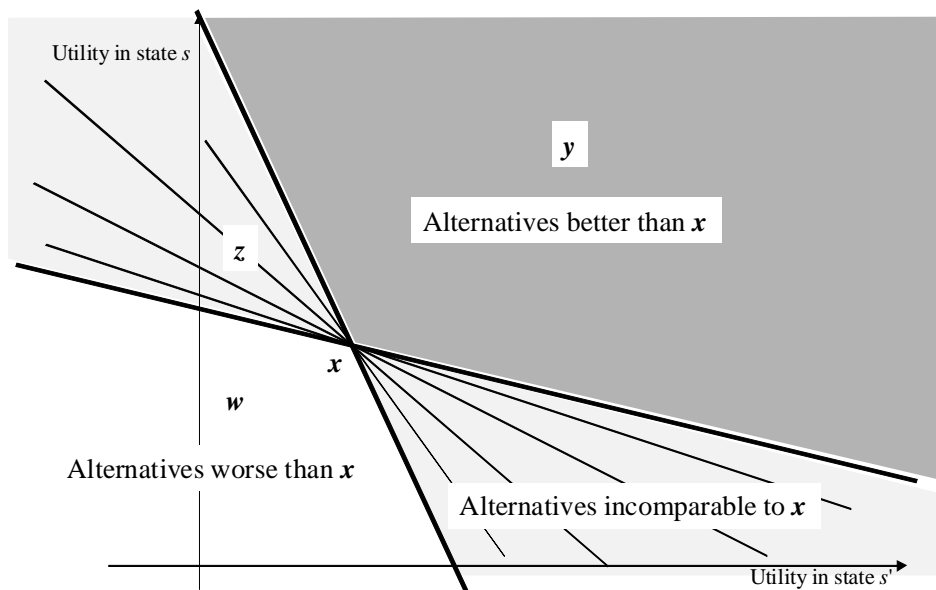


Figure 1: Incomplete Preferences

A bundle like y is preferred to x since it lies above all indifference lines through x . Similarly, x is preferred to w since w lies below all indifference lines through x . Finally, z is not comparable to x since it lies above some indifference lines through x and below others. Thus each bundle x generates three regions: bundles better than x , worse than x , and incomparable to x . This last set is empty only if there is a unique probability distribution over the two states (i.e. the preferences are complete).

Usual revealed preference arguments may not apply when preferences are incomplete. If y is chosen when x is available, we cannot say y is revealed preferred to x ; we can only say x is not revealed preferred to y . Choice among incomparable alternatives cannot be linked directly to preferences. This observation will have important implications for modeling implementation in mechanism design problems since it bears directly on the idea of incentive compatibility.

Bewley (1986) proposes a behavioral assumption, inertia, to deal with some choices among incomparable alternatives. The inertia assumption comprises two distinct assumptions: first, the existence of a distinguished alternative that can be considered the *status quo*, and second, that this status quo alternative is chosen as long as no feasible alternative is strictly preferred to it. Notice that this assumption is independent from the model of incomplete preferences described above. Indeed, we will make no assumptions regarding either the existence of a status quo alternative or inertia, in part because our focus is on robustness to agents' multiple priors. One difficulty with such an inertia assumption is identifying a plausible candidate for the role of status quo. In a mechanism design framework, a natural candidate for a status quo is the outside option. In this case, inertia could be used to refine the notion of individual rationality. This would not affect the meaning of incentive compatibility, however, since the status quo plays no role there.

3 The Setup

In this section, we describe a simple mechanism design framework with Knightian uncertainty, and introduce two notions of incentive compatibility. Our setup is entirely standard, except for the possibility that agents' perceptions of the state space reflect Knightian uncertainty. We focus on a framework with a single agent so that we can strip away issues pertaining to higher order beliefs and strategic uncertainty.

Let S be the set of states and O be the set of outcomes. We assume that S is a compact metric space. We let $\Delta(E)$ denote the set of all countably additive probability measures on a set E , and endow $\Delta(E)$ with the weak* topology. There is a single agent with privately known type $t \in T$. We restrict attention to direct mechanisms.¹¹ Any such mechanism is a function $\phi : T \times S \rightarrow O$ that specifies an outcome $\phi(\theta, s) \in O$ for any reported type $\theta \in T$ and any realized state $s \in S$. The agent's payoff function is

$$u : O \times T \times S \rightarrow \mathbf{R}.$$

If the agent reports θ while her true type is t , her *ex post utility* when the realized state is s is

$$u(\phi(\theta, s), t, s).$$

This framework can be modified to allow for many agents, by specifying that the type spaces of other agents be part of each agent's state space.

Each type $t \in T$ has a closed, convex set of beliefs $\Pi(t) \subset \Delta(S)$. We follow the standard assumption that the designer does not know the type of the agent, but does know the beliefs associated to each type. If the agent reports θ while her true type is t , we denote her expected payoff according to $\pi \in \Pi(t)$ by

$$E_{\pi} [u(\phi(\theta, s), t, s)].$$

In the standard Bayesian setup, $\Pi(t)$ is a singleton for each $t \in T$, and this expected value corresponds to the agent's *interim* expected utility. This case corresponds to the absence of Knightian uncertainty in our model.¹² We will refer to the opposite extreme where the set of beliefs of each type is the entire simplex, that is, $\Pi(t) = \Delta(S)$ for all $t \in T$, as "full ignorance".

A *mixed strategy* is a function $\sigma : T \rightarrow \Delta(T)$ specifying a probability distribution $\sigma(t) \in \Delta(T)$ for each $t \in T$.¹³ The expected payoff generated by $\sigma(t) \in \Delta(T)$ is denoted

$$E_{\sigma(t)} [u(\phi(\theta, s), t, s)],$$

¹¹In the appendix we verify that the revelation principle holds in our setting: for any equilibrium outcome of any (indirect) mechanism there exists a direct mechanism in which truth-telling is an equilibrium inducing the same outcome.

¹²The terms "ex post" and "interim" are in accordance with standard mechanism design terminology: *interim* refers to the stage at which the agent knows her type but has not observed the state, while *ex post* refers to the stage at which there is no uncertainty about the realized state.

¹³Since $\sigma(t)$ specifies a unique distribution for each type, the agent views the randomness induced by his use of a mixed strategy as risk rather than uncertainty.

and the interim expected utility of $\sigma(t)$ according to any $\pi \in \Pi(t)$ is

$$E_\pi [E_{\sigma(t)}[u(\phi(\theta, s), t, s)]] .$$

In this framework, there are several plausible definitions of incentive compatibility. The first is the standard notion of ex post incentive compatibility. Because it is an ex post concept, beliefs are irrelevant: truth-telling must be preferred to any misreport in each state.

Definition 1 *A mechanism $\phi : T \times S \rightarrow O$ is ex post incentive compatible for type t if for each $s \in S$*

$$u(\phi(t, s), t, s) \geq u(\phi(\theta, s), t, s) \quad \forall \theta \in T.$$

Matters are less straightforward at the interim stage. In settings without Knightian uncertainty, preferences are complete and there is no incentive to misreport if and only if reporting truthfully is (weakly) preferred to not doing so. In this case, interim incentive compatibility requires that truth telling yields (weakly) higher expected utility than any other strategy. This requirement is ambiguous when the agent's beliefs set is not a singleton. In particular, preferences are not complete and a mechanism may offer no incentive to misreport even if reporting truthfully is not preferred to misreporting: the two choices can be incomparable.

Two notions of interim incentive compatibility arise naturally in this context. The first notion requires that no strategy yields higher expected utility than truth-telling for all beliefs in the agent's belief set.

Definition 2 *A mechanism $\phi : T \times S \rightarrow O$ is maximal incentive compatible for type t if there exists no $\sigma(t) \in \Delta(T)$ such that*

$$E_\pi [E_{\sigma(t)} [u(\phi(\theta, s), t, s)]] > E_\pi [u(\phi(t, s), t, s)] \quad \forall \pi \in \Pi(t) .$$

The second notion of interim incentive compatibility requires that truth-telling yields (weakly) higher expected utility than misreporting for all beliefs in the agent's belief set.

Definition 3 *A mechanism $\phi : T \times S \rightarrow O$ is optimal incentive compatible for type t if for all $\sigma(t) \in \Delta(T)$*

$$E_\pi [u(\phi(t, s), t, s)] \geq E_\pi [E_{\sigma(t)} [u(\phi(\theta, s), t, s)]] \quad \forall \pi \in \Pi(t) .$$

We say that a mechanism is *ex post, maximal, or optimal incentive compatible* if it is ex post, maximal, or optimal incentive compatible for all $t \in T$.¹⁴ We focus on incentive compatibility, and disregard individual rationality, only to highlight that our main result does not depend on it.

¹⁴Note that to verify optimal incentive compatibility, it suffices to check for deviations in pure strategies, while for maximal incentive compatibility deviations in mixed strategies must also be checked.

On the other hand, following the same reasoning above, one would have two notions of individual rationality: *maximal individual rationality*, in which the outside option is not preferred to truth telling, and *optimal individual rationality*, in which truth telling is preferred to the outside option.

An optimal incentive compatible mechanism is also maximal incentive compatible while the converse does not necessarily hold. One can easily construct maximal incentive compatible mechanisms in which there exists a report that is not comparable to truthful reporting for some type.

The relationship between maximal and optimal incentive compatibility differs from the familiar relationship between best-response and strict best-response in the standard Bayesian framework because of the difference between incomparability and indifference.¹⁵ In particular, when preferences are strictly monotone, small changes in payoffs can always break ties due to indifference, but cannot in general eliminate incomparable alternatives.

The type space is $T=\{t,\theta\}$, the state space is $S=\{s,s'\}$, $\phi(\cdot)$ is a direct mechanism, and the agent's true type is t

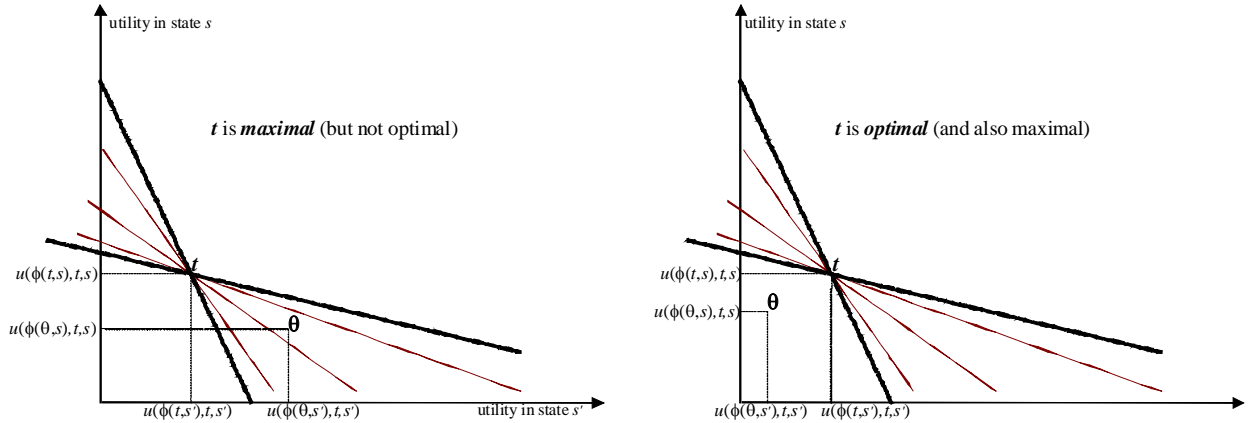


Figure 2: Maximal and Optimal Incentive Compatible Mechanism

Definitions 2 and 3 are illustrated by Figure 2. The axes display utility levels in an environment with only two states and two types. Given a direct mechanism ϕ , the point t represents the utility the agent gets from reporting truthfully while the point θ represents what she gets from misreporting. The graph on the left is a maximal mechanism: the agent's expected utility from telling the truth is sometimes larger and sometimes smaller than the expected utility from lying. The graph on the right is an optimal mechanism: the agent's expected utility from telling the truth is always larger than the expected utility from lying. Figure 2 also helps to illustrate that the interim incentive constraints become more restrictive as the amount of Knightian uncertainty perceived by the agent increases. As $\Pi(t)$ becomes larger, the maximal incentive constraints become less demanding while the optimal incentive constraints become tighter.

Another way to appreciate the distinction between maximal and optimal incentive compatibility comes from considering an agent's behavior in mechanisms that satisfy maximal or op-

¹⁵For example, indifference is transitive while incomparability need not be.

timal incentive compatibility. In the maximal case, it is difficult to conclude that the agent will report truthfully since, as in Figure 2, misreporting and truthtelling can be incomparable alternatives. Unlike indifference, this issue cannot be resolved by arbitrarily small payments or increases in utility. In contrast, if the mechanism makes truthtelling an optimal choice for the agent, then misreporting has (at least weakly) lower expected utility than truthtelling for *any* prior and for *any* possible misreport. A misreport can also be optimal in this case only if it yields the same expected utility as truthtelling for *every* prior. If beliefs are sufficiently rich, the only such misreports are those that are indifferent to truthtelling ex post, that is, such that $u(\phi(t, s), t, s) = E_\sigma[u(\phi(\theta, s), t, s)]$ for every state s . Similarly, a misreport that is not optimal can be maximal in this case only if it yields the same expected utility as truthtelling for some priors, and strictly lower expected utility for *all other* priors. In either case, just as with misreports that are indifferent to truthtelling in a standard Bayesian mechanism, truthtelling can be made strictly preferred by arbitrarily small payments or utility increases. Any other misreport must yield strictly lower expected utility for every prior. In this sense, optimal incentive compatibility reflects the requirement that the mechanism be robust to the presence of Knightian uncertainty.¹⁶

Optimal incentive compatibility can also be viewed as reflecting robustness to the agent’s attitude toward uncertainty. In this interpretation, rather than assuming that the agent has incomplete preferences represented as above and that this is known by the designer, the designer may identify a set of possible beliefs for the agent, but does not know the agent’s attitude toward uncertainty. Many different models of ambiguity are consistent with this framework, including maxmin expected utility, Choquet expected utility, the smooth ambiguity model and other models of second-order priors. This argument is similar to the separation between beliefs and ambiguity attitudes in Ghirardato, Maccheroni, and Marinacci (2004) and in Rigotti, Shannon, and Strzalecki (2008). Optimal incentive compatibility then corresponds to the requirement that the mechanism be robust to lack of detailed knowledge regarding how ambiguity is perceived by the agent.

Another view of our results takes a further step away from the interpretation of Knightian uncertainty, and instead follows the work on robustness to higher order beliefs in mechanism design, as a version of addressing the “Wilson critique”. In an influential paper, Bergemann and Morris (2005) argue that mechanisms should be robust to relaxing common knowledge assumptions among the planner and agents. They ask when ex post implementation is equivalent to interim implementation for every type space. This is interpreted as a reflection of robustness to the planner’s lack of knowledge of agents’ beliefs or higher order beliefs about other agents’ types. In the face of this potential lack of knowledge, interim implementation is required for every possible specification of agents’ beliefs over types. More specifically, they fix a payoff environment, which specifies the set of outcomes, the payoff type and utility function over outcomes and type profiles for each agent, and a social choice correspondence mapping payoff type profiles to sets of outcomes. Using an implicit representation of higher order beliefs as in Harsanyi (1967) and Mertens and Zamir (1985), types encode beliefs and the infinite hierarchy of higher order beliefs in their model. Thus fixing the payoff environment while varying agents’ beliefs over types generates different type spaces corresponding to the same underlying payoff types. Interim

¹⁶A similar reasoning applies if one considers individual rationality. Maximal individual rationality does not guarantee that the agent will participate in the mechanism while optimal individual rationality does (weakly) .

implementation for every type space then requires the mechanism to provide correct incentives for any possible specification of beliefs over types. Bergemann and Morris (2005) show that interim implementation implies ex post implementation provided the problem is separable, notably including quasilinear environments and problems in which the social choice correspondence is a function.

Formally, we consider a closely related notion of robustness, although our motivation is quite distinct. We also examine mechanisms that must provide correct incentives in a fixed payoff environment for a range of beliefs over states of nature, which we can take to include types of other agents in a multi-agent setting. By appealing to an implicit type space representation, our results could similarly be interpreted in terms of robustness to higher order beliefs. In contrast, our underlying foundation of uncertainty and incomplete preferences provides an equilibrium behavioral justification for the robustness notion we consider, as well as providing theoretical guidance for restricting the set of beliefs considered.

We restrict attention to social choice functions, rather than more general social choice correspondences. As Bergemann and Morris (2005) note, in this case considering direct mechanisms that use only information on payoff types is without loss of generality, and their main results hold with no further restrictions on the environment. In this case, their results can be viewed as embedding earlier results of Ledyard (1978) in the modern framework of type spaces. Allowing for extreme beliefs drives their conclusions: among all possible beliefs over types or states are those that are degenerate, assigning probability one to a given type profile or state. Interim implementation for the type space corresponding to these degenerate beliefs requires that the outcome of the mechanism is an equilibrium when this type profile or state is fixed, and since every degenerate belief is possible, interim implementation for all type spaces requires ex post implementation.¹⁷ For reference and comparison we give a version of this argument below for our general setting, in Lemma 1; see also Ledyard (1978). Our main results, however, are quite distinct. Instead we show that, surprisingly, in many common mechanism design settings, robustness to subtle and arbitrarily small variation in beliefs nonetheless requires ex post incentive compatible mechanisms. We discuss these points in more detail below following Theorem 1.

When $\Pi(t)$ is a singleton, maximal and optimal incentive compatibility collapse to the standard notion of interim incentive compatibility. At the other extreme case of full ignorance, in which the agent's belief set is the entire simplex, optimal incentive compatibility is equivalent to ex post incentive compatibility, while maximal incentive compatibility is equivalent to the property that truth-telling is not strictly dominated.¹⁸ We record these observations as Lemma 1.

¹⁷Bergemann and Morris (2005) also consider the restriction to full support beliefs. They show that their results carry over to interim implementation for all possible full support beliefs in the quasilinear case, or more generally, under compactness conditions, using a limiting argument as full support beliefs arbitrarily approximate degenerate distributions.

¹⁸Truth-telling is not strictly dominated for type t when there is no $\sigma(t) \in \Delta(T)$ such that

$$E_{\sigma(t)} [u(\phi(\theta, s), t, s)] > u(\phi(t, s), t, s) \quad \forall s \in S.$$

Lemma 1 *If $\Pi(t) = \Delta(S)$, then*

- (i) *a mechanism is optimal incentive compatible for type t if and only if it is ex post incentive compatible for type t ;*
- (ii) *a mechanism is maximal incentive compatible for type t if and only if truth-telling is not strictly dominated for type t .*

Proof (i) Ex post incentive compatibility always implies optimal incentive compatibility, so we need to prove only the other direction. Suppose that ϕ is optimal incentive compatible for type t but not ex post incentive compatible for that type. Then, there exists a state s and a report θ such that

$$u(\phi(t, s), t, s) < u(\phi(\theta, s), t, s).$$

Let π^s denote the measure assigning probability one to s . Since $\pi^s \in \Pi(t) = \Delta(S)$, truth-telling cannot be optimal incentive compatible.

(ii) Clearly if truth-telling is not strictly dominated, maximal incentive compatibility holds. For the converse, suppose that ϕ is maximal incentive compatible for type t . Then, there exists no $\sigma(t) \in \Delta(T)$ such that

$$E_{\pi} [u(\phi(t, s), t, s)] < E_{\pi} [E_{\sigma(t)} [u(\phi(\theta, s), t, s)]] \quad \forall \pi \in \Delta(S).$$

Let π^s denote the measure assigning probability one to some state s . Clearly, $\pi^s \in \Pi(t) = \Delta(S)$ for each state $s \in S$. For these measures, the above inequality becomes

$$u(\phi(t, s), t, s) < E_{\sigma(t)} [u(\phi(\theta, s), t, s)].$$

Since this applies to any $s \in S$, truth-telling is not strictly dominated. ■

Lemma 1 is related to some early work on implementation when the designer may not know the agents' beliefs in a standard Bayesian setting. Ledyard (1978, 1979) shows that a mechanism that is incentive compatible for any belief of the agent is equivalent to an ex post incentive compatible mechanism. As we discussed above, this framework is isomorphic to imposing optimal incentive compatibility under full ignorance. Lemma 1 translates in our framework a well known result in the mechanism design literature, thus is not surprising. In the next section, our main result shows that the equivalence between ex post incentive compatibility and optimal incentive compatibility goes beyond this simple case. Surprisingly, this equivalence can hold even when the agent's belief set is arbitrarily small.

4 Optimal and Ex Post Incentive Compatibility

In this section, we provide conditions under which optimal and ex post incentive compatibility are equivalent. Since ex post incentive compatibility always implies optimal incentive compatibility, we concentrate on establishing the reverse implication.

We begin by imposing regularity conditions on the payoff function that enable us to use the envelope theorem. These conditions are fairly common in the mechanism design literature.

Assumption 1 (a) *The type space is $T = [0, 1]$;*

(b) *The payoff function $u : O \times T \times S \rightarrow \mathbf{R}$ is differentiable with respect to t , and $u_2 := \frac{\partial u}{\partial t}$ is non-negative and bounded.*

Given the differentiability assumption, the non-negativity of u_2 is without additional loss of generality, since types can always be reordered. Assumption 1 does not rule out the possibility that u_2 equals zero on some interval. Therefore, two distinct types t and t' can have the same payoff function $u(\cdot, t, \cdot) = u(\cdot, t', \cdot)$ even if their belief sets are different.

We next provide a useful characterization of mechanisms that satisfy ex post incentive compatibility. Under Assumption 1, ex post incentive compatibility can be characterized in terms of an envelope condition and a cyclical monotonicity condition.

Definition 4 *A mechanism $\phi : T \times S \rightarrow O$ satisfies the ex post envelope condition if for each $s \in S$*

$$u(\phi(t', s), t', s) - u(\phi(t, s), t, s) = \int_t^{t'} u_2(\phi(\tau, s), \tau, s) d\tau \quad \forall t, t' \in T.$$

Definition 5 *A mechanism $\phi : T \times S \rightarrow O$ is ex post cyclically monotone if for all finite cycles $t_0, t_1, \dots, t_{N+1} = t_0$*

$$\sum_{k=0}^N [u(\phi(t_k, s), t_{k+1}, s) - u(\phi(t_k, s), t_k, s)] \leq 0$$

for each $s \in S$.

Rochet (1987) shows that in a quasilinear environment, cyclical monotonicity and ex post incentive compatibility are equivalent. This result is extended by Bikhchandani, Chatterji, Lavi, Mu'alem, Nisan, and Sen (2006), in which cyclical monotonicity is replaced by the less stringent weak monotonicity, and a richness condition is imposed on the type space. Since our environment is not quasilinear, we cannot appeal directly to these results. The next lemma thus extends Rochet's result to our setting. Notice that cyclical monotonicity and weak monotonicity reduce to the more familiar ex post monotonicity in the standard auction setting (see, for example, Krishna (2002)).

Lemma 2 *Suppose Assumption 1 holds. Then a mechanism is ex post incentive compatible if and only if it satisfies ex post cyclical monotonicity and the ex post envelope condition.*

Proof Fix $s \in S$, and set $w(\theta, t) := u(\phi(\theta, s), t, s)$. Using this notation, the ex post envelope condition, ex post monotonicity, and ex post incentive compatibility can be rewritten as follows:

$$w(t', t') - w(t, t) = \int_t^{t'} w_2(\tau, \tau) d\tau \quad \forall t, t' \in T, \quad (\text{EPEC})$$

$$\sum_{k=0}^N [w(t_k, t_{k+1}) - w(t_k, t_k)] \leq 0 \quad \text{for all finite cycles } t_0, t_1, \dots, t_{N+1} = t_0 \quad (\text{EPCM})$$

and

$$w(t, t) \geq w(\theta, t) \quad \forall t, \theta \in T. \quad (\text{EPIC})$$

where $w_2(\cdot)$ denotes the derivative with respect to the second argument.

First we show that EPEC and EPCM imply EPIC. To that end, take $t_0 \in [0, 1]$ to be arbitrary, and define the function $V(t)$ by

$$V(t) := \sup_{\{\text{all chains } t_0 \text{ to } t_{N+1}=t\}} \sum_{k=0}^N [w(t_k, t_{k+1}) - w(t_k, t_k)]$$

By definition, $V(t_0) = 0$, and

$$V(t_0) \geq V(t) + w(t, t_0) - w(t, t)$$

which implies that $V(t)$ is finite for all $t \in [0, 1]$. Then as in Rochet (1987), cyclical monotonicity implies

$$V(t) \geq V(t') + w(t', t) - w(t', t') \quad \forall t, t' \in T. \quad (1)$$

This can be rewritten as

$$V(t) - V(t') \geq w(t', t) - w(t', t')$$

Switching the arguments above yields

$$V(t') - V(t) \geq w(t, t') - w(t, t).$$

Combining the two previous inequalities yields

$$w(t, t) - w(t, t') \geq V(t) - V(t') \geq w(t', t) - w(t', t').$$

Taking $t > t'$ and dividing by $t - t'$ yields

$$\frac{w(t, t) - w(t, t')}{t - t'} \geq \frac{V(t) - V(t')}{t - t'} \geq \frac{w(t', t) - w(t', t')}{t - t'}.$$

Taking the limit as t' goes to t , we conclude

$$w_2(t, t) = V'(t) \quad \forall t \in (0, 1).$$

Together with EPEC this implies

$$V(t) = w(t, t) + k$$

for some constant k . Hence, equation 1 can be rewritten

$$w(t, t) \geq w(t', t') + w(t', t) - w(t', t') \quad \forall t, t' \in T;$$

or

$$w(t, t) \geq w(t', t) \quad \forall t, t' \in T,$$

which is EPIC.

For the converse, suppose that EPIC holds. Using Assumption 1, EPEC follows immediately from the envelope theorem as in Theorem 2 in Milgrom and Segal (2002). Now consider any finite cycle $t_0, t_1, \dots, t_{N+1} = t_0$. Incentive compatibility implies

$$\forall k = 0, \dots, N, \quad w(t_k, t_{k+1}) - w(t_k, t_k) \leq 0.$$

Summing all these inequalities yields

$$\sum_{k=0}^N [w(t_k, t_{k+1}) - w(t_k, t_k)] \leq 0,$$

or

$$\sum_{k=0}^N [w(t_{k+1}, t_k) - w(t_k, t_k)] \leq 0,$$

which implies EPCM. ■

The second set of assumptions needed for the main result pertain to the agent's beliefs. Roughly, beliefs are required to be sufficiently rich and locally overlapping as types vary. We formalize these ideas as follows.

Definition 6 A set $\Pi \subset \Delta(S)$ has full dimension if, given any continuous function $g : S \rightarrow \mathbf{R}$,

$$\int_S g(s) d\pi = 0 \quad \forall \pi \in \Pi \quad \text{implies } g = 0.$$

Definition 7 The agent's beliefs are fully overlapping if for each $t \in T$ there exists a neighborhood $N(t) \subset T$ such that $\bigcap_{t' \in N(t)} \Pi(t')$ has full dimension.

At the end of this section, we elaborate on how this assumption can be interpreted as a consequence of a form of continuity of the mapping of types into belief sets. We now turn to the main result.

Theorem 1 Suppose that Assumption 1 holds and that the agent's beliefs are fully overlapping. Then any mechanism that is optimal incentive compatible and ex post cyclically monotone is also ex post incentive compatible.

Proof Lemma 2 has established that ex post cyclical monotonicity is necessary for ex post incentive compatibility. The main step in the proof below then consists in using the envelope theorem in integral form (Milgrom and Segal (2002)) and the assumption of fully overlapping beliefs to arrive at the ex post envelope condition. The result then follows immediately from Lemma 2.

Let ϕ be a mechanism that satisfies optimal incentive compatibility. By definition, for all $t, \theta \in T$,

$$\int_S u(\phi(t, s), t, s) d\pi - \int_S u(\phi(\theta, s), t, s) d\pi \geq 0 \quad \forall \pi \in \Pi(t).$$

Fix $t_0 \in T$ and, using fully overlapping beliefs, choose a neighborhood $N(t_0) \subset T$ of t_0 such that $\Pi^N(t_0) := \bigcap_{t' \in N(t_0)} \Pi(t')$ has full dimension. For each $\pi \in \Pi^N(t_0)$ and each $t' \in N(t_0)$, the inequalities above together with an envelope theorem (see Theorem 2 in Milgrom and Segal (2002)) ensure that

$$\int_S u(\phi(t_0, s), t_0, s) d\pi - \int_S u(\phi(t', s), t', s) d\pi = \int_{t'}^{t_0} \left(\int_S u_2(\phi(\tau, s), \tau, s) d\pi \right) d\tau \quad (2)$$

or equivalently

$$\int_S \left[u(\phi(t_0, s), t_0, s) - u(\phi(t', s), t', s) - \int_{t'}^{t_0} u_2(\phi(\tau, s), \tau, s) d\tau \right] d\pi = 0 \quad \forall \pi \in \Pi^N(t_0).$$

Since $\Pi^N(t_0)$ has full dimension, for each $t' \in N(t_0)$,

$$u(\phi(t_0, s), t_0, s) - u(\phi(t', s), t', s) = \int_{t'}^{t_0} u_2(\phi(\tau, s), \tau, s) d\tau \quad \forall s \in S. \quad (3)$$

This in turn implies that for almost all $t' \in N(t_0)$,

$$\frac{\partial}{\partial t} u(\phi(t', s), t', s) = u_2(\phi(t', s), t', s) \quad \forall s \in S.$$

Since t_0 was arbitrary, we conclude that equation (3) holds for all $t, t' \in T$. Thus ϕ satisfies the ex post envelope condition. By assumption, ϕ is also ex post monotone, so Lemma 2 yields the desired conclusion that ϕ is ex post incentive compatible. \blacksquare

Assumption 1 is indispensable for this equivalence result. In particular, if T is finite, even if all types have the same belief set of full dimension, a large class of mechanisms will be ex post cyclically monotone and optimal incentive compatible but not ex post incentive compatible. Intuitively, this is because with a finite type space there is ‘slackness’ in the incentive compatibility constraints. The following example shows that the fully overlapping beliefs condition is also critical.

Example 1 Let $S = \{1, 2\}$, and assume the agent has quasilinear utility with value $t \in [0, 1]$ for a given object. An outcome is determined by the pair (q, m) , where q denotes the probability that the agent is awarded the object and m denotes her payment. A mechanism specifies an outcome $q(\theta, s)$, $m(\theta, s)$, for any pair $(\theta, s) \in [0, 1] \times \{1, 2\}$. The agent’s payoff function is $u(\phi(\theta, s), t, s) = tq(\theta, s) - m(\theta, s)$. A probability $\pi \in \Delta(S)$ is identified by the number $\pi_1 := \Pr[s = 1]$, while a belief set $\Pi(t)$ corresponds to an interval $[\underline{\pi}_1(t), \bar{\pi}_1(t)] \subset [0, 1]$. Let $\varepsilon \in (0, \frac{1}{6})$, and

$$[\underline{\pi}_1(t), \bar{\pi}_1(t)] = \begin{cases} [\frac{1}{3} - 2\varepsilon, \frac{1}{3} - \varepsilon], & 0 \leq t \leq \frac{1}{2}; \\ [\frac{2}{3} + \varepsilon, \frac{2}{3} + 2\varepsilon], & \frac{1}{2} < t \leq 1. \end{cases}$$

Consider the mechanism (q, m) , where $q(\theta, s) = 1$ for all (θ, s) , $g > 0$, and

$$m(\theta, s) = \begin{cases} \frac{1}{2}g, & \text{if } \theta > \frac{1}{2} \text{ and } s = 1; \\ -g, & \text{if } \theta > \frac{1}{2} \text{ and } s = 2; \\ 0, & \text{if } \theta \leq \frac{1}{2}. \end{cases}$$

Since $\pi(t - \frac{1}{2}g) + (1 - \pi)(t + g) = t + g(1 - \frac{3}{2}\pi)$, the interim expected utility function

$$U(\phi(\theta, s), t) = \begin{cases} t + g(1 - \frac{3}{2}\pi) & \text{if } \theta > \frac{1}{2}, \\ t & \text{if } \theta \leq \frac{1}{2}, \end{cases}$$

is maximized by any $\theta > \frac{1}{2}$ if $\pi < \frac{2}{3}$, and by any $\theta \leq \frac{1}{2}$ if $\pi > \frac{2}{3}$. Since $\pi < \frac{2}{3}$ for all $\pi \in \Pi(t)$ when $t \leq \frac{1}{2}$, and $\pi > \frac{2}{3}$ for all $\pi \in \Pi(t)$ when $t > \frac{1}{2}$, the mechanism satisfies optimal incentive compatibility. However it is easy to see that ϕ is not ex post incentive compatible: since $g > 0$ any type $t > \frac{1}{2}$ prefers to report $\theta \leq \frac{1}{2}$ when $s = 2$.

Lemma 2 and Theorem 1 can be generalized to multi-dimensional type spaces. For example, suppose that (i) the outcome space consists of all probability distribution over a finite set, i.e. $O = \Delta(\{1, \dots, k\})$, so that any mechanism can be represented by $\phi(\theta, s) = (\phi_1(\theta, s), \dots, \phi_k(\theta, s))$, where $\phi_j(\theta, s)$ denotes the probability of choosing outcome j ; (ii) T is a smoothly connected subset of \mathbf{R}^k , i.e. for any two points $t, t' \in T$ there exists a differentiable function $f : [0, 1] \rightarrow T$ such that $f(0) = t$ and $f(1) = t'$; and (iii) $u(\phi(\theta, s), s, t) = \sum_{j=1}^k t_j \phi_j(\theta, s)$. In this case, the function $\tilde{u}(j, s, t) = f(t_j(s))$ satisfies Assumption 1, and the proof of Theorem 1 follows, with the requisite change that the integrals in equations (2) and (3) are interpreted as path integrals along any path connecting t and t' .

Theorem 1 shows that the presence of Knightian uncertainty, when taken together with the particular notion of robustness imposed by optimal incentive compatibility, can severely limit the set of feasible mechanisms. In particular, the designer can only choose among ex post incentive compatible mechanisms. This result is obtained despite maintaining the standard common knowledge assumptions that make the problem easier to solve for the designer. Nonetheless, this informational advantage is not sufficient to utilize mechanisms that are not ex post incentive compatible.

Our results are complementary to recent work on robust mechanism design, as we discussed above. Bergemann and Morris (2005) consider the problem of implementing a given social choice correspondence. In their setting, agent i has payoff function $u_i : Y \times \Theta_i \rightarrow \mathbf{R}$, where Y denotes the set of feasible outcomes and Θ_i denotes the set of all possible ‘‘payoff types’’. The agent’s privately known type $t_i \in T_i$ determines both her payoff type θ_i , via a function $\hat{\theta}_i : T_i \rightarrow \Theta_i$, and her belief about the other agents’ types $\hat{\pi}_i(\cdot | t_i) \in \Delta(T_{-i})$. A type space in this environment is a collection $(T_i, \hat{\theta}_i, \hat{\pi}_i)_{i=1}^I$. A *payoff type space* is a type space in which for each i , $T_i = \Theta_i$ and $\hat{\theta}_i$ is the identity map. Our setup can be interpreted as a payoff type space where the state s includes the type profile of other agents whose behavior is not modeled explicitly.

Bergemann and Morris (2005) show that, with quasi-linear payoff functions and unrestricted payments, ex post incentive compatibility is equivalent to interim incentive compatibility in all payoff type spaces in which the agents' beliefs are consistent with a common prior. The key observation, which can be viewed as embedding earlier results of Ledyard (1978) in the modern framework of type spaces, is the following. If a mechanism ϕ satisfies interim incentive compatibility for all distributions, then it must also satisfy interim incentive compatibility in the type space where all agents but i have a fixed payoff type profile θ_{-i} . Since this is true for any profile $\theta_{-i} \in \Theta_{-i}$, the mechanism ϕ must be ex post incentive compatible. This reasoning does not rely on the cardinality of the type space and can be used without essential alterations to establish the equivalence between optimal and ex post incentive compatibility in our setting for the case in which $\Pi(t) = \Delta(S)$ for all $t \in T$. We have applied a similar reasoning in the proof of Lemma 1. In contrast, Theorem 1 shows that with a continuum of types, the equivalence between optimal incentive compatibility and ex post incentive compatibility may hold even when the belief sets of all types are significantly restricted.

We close this section by examining the assumption that beliefs are fully overlapping. A sufficient condition for this assumption to be satisfied is that the correspondence mapping types into belief sets is in some sense continuous. As the following example makes clear, this restriction can be satisfied by sets of beliefs that are small relative to the simplex. The latter observation is important when interpreting our result in light of the recent literature on robustness in mechanism design.

A sufficient condition for the belief set to have full dimension comes from the observation that a convex set $\Pi \subset \Delta(S)$ has full dimension whenever its algebraic interior in $\Delta(S)$ is non-empty, where the algebraic interior of Π is given by

$$\text{alg-int } \Pi := \{\pi \in \Pi : \forall \tilde{\pi} \in \Delta(S) \text{ there exists } \delta \in (0, 1] \text{ such that } (1 - \delta)\pi + \delta\tilde{\pi} \in \Pi\}.$$

A stronger sufficient condition, equivalent when S is finite, is that Π has non-empty relative interior. Using this observation, a simple example of such a set with full dimension arises from the common ε -contamination model of ambiguity, in which for fixed $\varepsilon \in (0, 1)$ and $\bar{\pi} \in \Delta(S)$, the belief set is given by

$$\Pi_\varepsilon := \{\pi \in \Delta(S) : \pi = (1 - \varepsilon)\bar{\pi} + \varepsilon\tilde{\pi} \text{ for some } \tilde{\pi} \in \Delta(S)\}.$$

The set Π_ε has full dimension for any $\varepsilon \in (0, 1)$.

Beliefs are fully overlapping under a mild form of continuity of the correspondence describing the agent's beliefs as her type varies. This point is formalized by the following result.

Theorem 2 *If the correspondence $\Pi : T \rightarrow 2^{\Delta(S)}$ is lower hemi-continuous and $\Pi(t)$ has non-empty relative interior for each $t \in T$, then the beliefs $\{\Pi(t) : t \in T\}$ are fully overlapping.*

Proof For each t , let $\text{rint}\Pi(t)$ denote the relative interior of $\Pi(t)$, which is non-empty by assumption. The correspondence $t \mapsto \text{rint}\Pi(t)$ is lower hemi-continuous, so has a continuous selection $\pi(t) \in \text{rint}\Pi(t)$ for each t . Fix t , and choose $\epsilon > 0$ and a neighborhood $N(t)$ such

that $B_\epsilon(\pi(t')) \subset \text{rint}\Pi(t')$ for each $t' \in N(t)$. Next, using the continuity of $t \mapsto \pi(t)$, choose a neighborhood $\hat{N}(t) \subset N(t)$ such that for all $t' \in \hat{N}(t)$, $\pi(t') \in B_{\epsilon/2}(\pi(t))$. Then by construction, $B_{\epsilon/2}(\pi(t)) \subset B_\epsilon(\pi(t')) \subset \text{rint}\Pi(t')$ for each $t' \in \hat{N}(t)$, hence $B_{\epsilon/2}(\pi(t)) \subset \bigcap_{t' \in \hat{N}(t)} \text{rint}\Pi(t')$. ■

Returning to the ϵ -contamination model for an example, if $t \mapsto (\bar{\pi}(t), \epsilon(t)) \in \Delta(S) \times (0, 1)$ is continuous, where $\{\bar{\pi}(t) : t \in T\}$ are mutually absolutely continuous, and $\Pi(t) := \Pi_{\epsilon(t)}(\bar{\pi}(t))$, then beliefs are fully overlapping. Notice that as ϵ goes to zero this model converges to the standard model in which Knightian uncertainty is ruled out. In that case, we know that the class of interim incentive compatible mechanisms is much larger than the class of ex post incentive compatible mechanisms. Yet under arbitrarily small amount of Knightian uncertainty, optimal incentive compatibility and ex post incentive compatibility can coincide.

5 Full Extraction in Knightian Mechanisms

In this section we examine the extent to which private information can generate rents in the presence of Knightian uncertainty. The answer to this question depends on the notion of incentive compatibility used and on how information rents are defined in the presence of Knightian uncertainty. We already know from Theorem 1 that, with a continuum of types, optimal incentive compatibility is equivalent to ex post incentive compatibility under some conditions on beliefs. Whenever this result applies, full extraction with ex post incentive compatible and ex post individually rational mechanisms is not feasible. For this reason we focus on the case with finitely many types.

We adopt a single agent setup similar to McAfee and Reny (1992). The agent can participate in a game that will leave her with ex post rents. These rents depend on her private information, summarized by a set of types T , and on publicly observed information, summarized by a set of states S . We focus on the case where both the type space T and the state space S are finite and, with slight abuse of notation, we use S and T to denote both the sets and their cardinality. The precise nature of the game (an auction, a bargaining game, etc.) is irrelevant for our purposes. We take the agent's payoff function $r : T \times S \rightarrow \mathbf{R}_+$ as a primitive, with $r(t) : S \rightarrow \mathbf{R}_+$ denoting the rents for type t as function of the publicly observed state s .

The designer (a seller, a mediator, etc.) can charge the agent for participating in the game, while the agent can choose whether or not to participate; if she does not participate her payoff is zero. Since the realization of s is publicly observable, the designer can charge a participation fee contingent on each realized state. Let $z \in \mathbf{R}^S$ denote a *participation fee schedule*. The designer can offer a finite menu $Z \subset \mathbf{R}^S$ of participation fee schedules from which the agent selects one.

Since the agent perceives Knightian uncertainty about the state space S , we assume that for each $t \in T$ there is a closed and convex set $\Pi(t) \subset \Delta(S)$ describing the beliefs of type t . The extent to which the designer can construct the menu Z to extract the agent's rent depends on the properties of the agent's belief sets and on the notion of incentive compatibility invoked. We explore characterizations of rent extraction in this setup, first under maximal incentive compatibility and then under the more restrictive notion of optimal incentive compatibility. In the last

subsection we discuss the robustness of the necessary and sufficient conditions for full extraction.

5.1 Surplus extraction with maximal choices

We begin by thinking of incentive compatibility in terms of maximal choices. For a given menu Z , this means that the agent's choice from Z should be "minimal" in terms of expected payments: there is no alternative which has lower expected cost for all beliefs.

Definition 8 *A participation fee schedule $z^m \in Z$ is minimal for type t in a menu Z if there is no mixed strategy $\sigma \in \Delta(Z)$ such that*

$$\sum_{z \in Z} \sigma_z [\pi \cdot z] < \pi \cdot z^m \quad \forall \pi \in \Pi(t).$$

When the agent's choices are minimal, the corresponding notion of surplus for type $t \in T$, relative to a given menu Z , is

$$S^m(t) := \{\pi \cdot [r(t) - z] : z \text{ is minimal for type } t \text{ in } Z, \pi \in \Pi(t)\}.$$

Note that S^m also depends on the rent function r and the menu Z , although we have suppressed this dependence.

Corresponding to the notion of minimal choices, we have the following definition of full rent extraction.

Definition 9 *A menu of participation fee schedules Z achieves maximal full extraction of the rent function r if for each $t \in T$ there exists $z(t) \in Z$ that is minimal for type t in Z and for which*

$$\pi \cdot [r(t) - z(t)] = 0 \text{ for some } \pi \in \Pi(t).$$

The designer can achieve maximal full extraction if for every $r \in \mathbf{R}_+^{S \times T}$ there exists a menu Z that achieves maximal full extraction of r .

Using the notation defined above, maximal rent extraction of a given rent function r can also be described as finding a menu Z for which $0 \in S^m(t)$ for each $t \in T$.

Under maximal incentive compatibility, as Knightian uncertainty increases (the set $\Pi(t)$ becomes larger) the incentive constraints become weaker. Maximal full rent extraction is achievable if there exists a *selection* from the correspondence $\{\Pi(t) : t \in T\}$ that satisfies the correlation condition familiar from the work of Crémer and McLean (1985).

Theorem 3 *The designer can achieve maximal full rent extraction if there exists a selection $\{\pi(t) : \pi(t) \in \Pi(t) \text{ for each } t \in T\}$ such that*

$$\forall t \in T : \quad \pi(t) \notin \text{co} \{ \{ \pi(t') \}_{t' \neq t} \} \tag{MFE}$$

where $\text{co } A$ denotes the convex hull of the set A .

Proof Fix a rent function $r \in \mathbf{R}_+^{T \times S}$. Suppose that (MFE) holds, and choose a selection $\{\pi(t) : \pi(t) \in \Pi(t) \text{ for each } t \in T\}$ such that

$$\forall t \in T : \quad \pi(t) \notin \text{co} \{ \{ \pi(t') \}_{t' \neq t} \}.$$

For each t , using the Separating Hyperplane Theorem, choose $\tilde{z}_t \in \mathbf{R}^S$ such that

$$\begin{aligned} \pi(t) \cdot \tilde{z}_t &= 0 \\ \pi(t') \cdot \tilde{z}_t &> 0 \quad \forall t' \neq t, \end{aligned}$$

and define

$$c_t := \pi(t) \cdot r(t)$$

By construction, for every $\alpha > 0$,

$$\pi(t) \cdot r(t) - c_t - \alpha \pi(t) \cdot \tilde{z}_t = 0$$

Then choose scaling factors $\alpha_t > 0$ for each t so that $\forall t' \neq t$,

$$\pi(t) \cdot r(t) - c_t - \alpha_t \pi(t) \cdot \tilde{z}_t \geq \pi(t') \cdot r(t') - c_{t'} - \alpha_{t'} \pi(t') \cdot \tilde{z}_{t'}$$

Because $\pi(t) \cdot \tilde{z}_{t'} > 0$ while $\pi(t') \cdot \tilde{z}_t = 0$, such a collection $\{\alpha_t\}_{t \in T}$ exists.

Also, for each t , set

$$z_t := c_t + \alpha_t \tilde{z}_t$$

and define the menu $Z := \{z_t : t \in T\}$. For all $t \in T$ we have

$$\pi(t) \cdot [r(t) - z_t] \geq \pi(t) \cdot [r(t) - z_{t'}] \quad \forall t' \neq t,$$

which guarantees that z_t is minimal for type t in the menu Z . (Note that it suffices to consider pure strategies because all comparisons are made with the same distribution $\pi(t)$); and by construction, for each $t \in T$

$$\pi(t) \cdot [r(t) - z_t] = 0 \in S^m(t)$$

Thus the menu Z achieves maximal full rent extraction of the rent function r . ■

5.2 Surplus extraction with optimal choices

When incentive compatibility is defined in terms of optimal choices, the presence of ambiguity strengthens the incentive constraints, and full rent extraction becomes more demanding. In this case, we require the agent's choice from the proposed menu of participation fee schedules Z to be “pessimal” with respect to expected payments: all alternatives have higher expected cost for all beliefs.

Definition 10 *A participation fee schedule z^p is pessimal for type t in a menu Z if for all mixed strategies $\sigma \in \Delta(Z)$*

$$\pi \cdot z^p \leq \sum_{z \in Z} \sigma_z [\pi \cdot z] \quad \forall \pi \in \Pi(t).$$

In analogy with the case of minimal choices, we define the corresponding measure of surplus

$$S^o(t) := \{\pi \cdot [r(t) - z] : z \text{ is pessimal for type } t \text{ in } Z, \pi \in \Pi(t)\}$$

and the corresponding notion of full rent extraction.

Definition 11 *The menu Z achieves optimal full rent extraction of the rent function r if for each $t \in T$:*

$$\pi \cdot [r(t) - z(t)] \geq 0 \text{ for all } z(t) \text{ pessimal for } t \text{ in the menu } Z \text{ and for all } \pi \in \Pi(t)$$

with equality for at least one such $z(t)$ and $\pi \in \Pi(t)$.

The designer can achieve optimal full rent extraction if for every rent function $r \in \mathbf{R}_+^{S \times T}$ there exists a menu Z that achieves optimal full rent extraction of r .

Optimal full rent extraction of a given rent function r can alternatively be characterized via the surplus measure for each $t \in T$ as the requirement:

$$0 \in S^o(t) \subset [0, a(t)], \text{ where } a(t) \geq 0$$

Next we provide necessary and sufficient conditions for optimal full rent extraction. The sufficient condition is a uniform version of the selection condition that ensures maximal full extraction, and requires that beliefs be sufficiently different across types. The necessary condition we give is weaker; it requires that beliefs not be too similar across types.¹⁹ We formalize this below.

Theorem 4 *The designer can achieve optimal full rent extraction if*

$$\forall t \in T : \quad \Pi(t) \cap \text{co} \{ \cup_{t' \neq t} \Pi(t') \} = \emptyset \quad (\text{OFE})$$

Moreover, the designer can achieve optimal full rent extraction only if

$$\forall t \in T : \quad \Pi(t) \not\subset \text{co} \{ \cup_{t' \neq t} \Pi(t') \} \quad (\text{NOFE})$$

Proof First, we claim that the condition (OFE) is sufficient for optimal full rent extraction. To see this, fix $r \in \mathbf{R}_+^{S \times T}$, and suppose that (OFE) holds. For each t , using the Separating Hyperplane Theorem, choose $\tilde{z}_t \in \mathbf{R}^S$ such that

$$\begin{aligned} \pi(t) \cdot \tilde{z}_t &\leq 0 & \forall \pi(t) \in \Pi(t) \\ \pi(t') \cdot \tilde{z}_t &> 0 & \forall \pi(t') \in \Pi(t'), \forall t' \neq t \end{aligned}$$

¹⁹The gap between the necessary and sufficient conditions arises due to the slack in the incentive compatibility constraints for some beliefs.

Adjusting \tilde{z}_t if necessary, \tilde{z}_t can be chosen such that in addition

$$\max_{\pi(t) \in \Pi(t)} \pi(t) \cdot \tilde{z}_t = 0$$

For each t , set

$$c_t := \min_{\pi(t) \in \Pi(t)} \pi(t) \cdot r(t) \quad \text{s.t.} \quad \pi(t) \cdot \tilde{z}_t = 0$$

By construction, for every $\alpha > 0$,

$$\begin{aligned} \pi(t) \cdot r(t) - c_t - \alpha \pi(t) \cdot \tilde{z}_t &\geq 0 & \forall \pi(t) \in \Pi(t) \\ \pi(t) \cdot r(t) - c_t - \alpha \pi(t) \cdot \tilde{z}_t &= 0 & \text{for some } \pi(t) \in \Pi(t) \end{aligned}$$

Then for each t , choose scaling factors $\alpha_t > 0$ so that $\forall t' \neq t$,

$$\pi(t) \cdot r(t) - c_t - \alpha_t \pi(t) \cdot \tilde{z}_t \geq \pi(t) \cdot r(t) - c_{t'} - \alpha_{t'} \pi(t) \cdot \tilde{z}_{t'} \quad \forall \pi(t) \in \Pi(t)$$

Because $\pi(t) \cdot \tilde{z}_{t'} > 0 \forall \pi(t) \in \Pi(t)$ while $\pi(t) \cdot \tilde{z}_t \leq 0 \forall \pi(t) \in \Pi(t)$, this is possible.

For each t , set

$$z_t := c_t + \alpha_t \tilde{z}_t$$

and define the menu $Z := \{z_t : t \in T\}$. For each $t \in T$:

$$\pi \cdot [r(t) - z_t] \geq 0 \quad \forall \pi \in \Pi(t)$$

with equality for some $\pi(t) \in \Pi(t)$. Thus the menu Z achieves optimal full rent extraction of r .

To see that (NOFE) is necessary for optimal full rent extraction, fix $t_0 \in T$ and consider the rent function $r(t) = (t - t_0)^2$. Suppose Z is a menu that achieves optimal extraction of the rents given by r .

For each $t \neq t_0$, choose $z(t) \in Z$ pessimal for t in Z and $\pi(t) \in \Pi(t)$ such that

$$\pi(t) \cdot [r(t) - z(t)] = 0$$

Similarly, let $z(t_0) \in Z$ be pessimal for t_0 in Z .

Suppose there exists $\mu \in \Delta(T \setminus \{t_0\})$ such that

$$\pi(t_0) := \sum_{t \neq t_0} \mu_t \pi(t) \in \Pi(t_0)$$

Then

$$\begin{aligned}
0 = \pi(t_0) \cdot r(t_0) &\geq \pi(t_0) \cdot z(t_0) \\
&= \sum_{t \neq t_0} \mu_t \pi(t) \cdot z(t_0) \\
&\geq \sum_{t \neq t_0} \mu_t \pi(t) \cdot z(t) \\
&= \sum_{t \neq t_0} \mu_t \pi(t) \cdot r(t) \\
&= \sum_{t \neq t_0} \mu_t (t - t_0)^2 \\
&> 0
\end{aligned}$$

But this is impossible. ■

5.3 Genericity of optimal full extraction

We conclude this section with results investigating the robustness of the necessary and sufficient conditions for full extraction identified above. The work of Crémer and McLean and others uncovered the connection between correlated beliefs and full extraction in standard Bayesian mechanism design. Perhaps the most powerful and negative aspect of this work was showing that these conditions are generic in an appropriate sense. Related work showed that these generic conditions lead to full extraction in a wide array of settings with private information. We seek a similar measure of the extent of full rent extraction and the existence of information rents in the presence of Knightian uncertainty.

To formalize this discussion, suppose $|T| \geq |S|$, so there are at least as many types as states. Recall that in the standard Bayesian setting, each type $t \in T$ is associated with a unique conditional distribution $\pi(t) \in \Delta(S)$, and full rent extraction is possible if and only if the correlation condition (MFE) holds for the collection $\{\pi(t) : t \in T\}$. It is straightforward to see that the subset of $\Delta(S)^T$ on which this condition is satisfied is an open set of full Lebesgue measure. In a setting with Knightian uncertainty, each type t holds a set of distributions $\Pi(t)$ drawn from

$$\mathcal{C} := \{\Pi \subset \Delta(S) : \Pi \text{ is closed and convex}\}$$

and full extraction is characterized in terms of conditions on the collection $\{\Pi(t) : t \in T\} \in \mathcal{C}^T$. To gauge how widespread the absence of information rents is in this setting, we seek to measure the size of the subset of \mathcal{C}^T corresponding to the various conditions we have identified above.

To make this precise, we endow \mathcal{C} with the Hausdorff topology, and \mathcal{C}^T with the product topology. Let \mathcal{M} denote the subset of \mathcal{C}^T satisfying (MFE), \mathcal{O} denote the subset of \mathcal{C}^T satisfying (OFE), and \mathcal{N} denote the subset of \mathcal{C}^T violating (NOFE). Thus \mathcal{M} is a set on which maximal full extraction is always possible. Similarly, \mathcal{O} is a set of beliefs for which optimal full extraction is always possible, and \mathcal{N} is a set for which optimal full extraction is never possible.

The set \mathcal{C}^T is infinite-dimensional, which means the issue of measuring the sizes of these sets is no longer straightforward due to the absence of a natural analogue of Lebesgue measure in infinite-dimensional spaces. Genericity in these cases is typically defined either using topological notions, such as open and dense or residual, or using measure-theoretic notions such as prevalence. Prevalence and its complement, shyness, developed by Christensen (1974) and Hunt, Sauer, and Yorke (1992), and made relative by Anderson and Zame (2001), are analogues of Lebesgue measure 0 and full Lebesgue measure that more closely mimic properties of Lebesgue measure in many problems.²⁰ We first give formal definitions, and then discuss some important properties shared by these notions of genericity.

Because we are interested in the relative size of subsets of \mathcal{C}^T , we use the relative notions of prevalence and shyness developed by Anderson and Zame (2001) for use in a convex subset which may be a shy subset of the ambient space. The formal definitions are given below.

Definition 12 *Let Z be a topological vector space and let $C \subset Z$ be a convex Borel subset of Z which is completely metrizable in the relative topology. Let $c \in C$. A universally measurable subset $E \subset Z$ is shy in C at c if for each $\delta > 0$ and each neighborhood W of 0 in Z , there is a regular Borel probability measure μ on Z with compact support such that $\text{supp } \mu \subset (\delta(C - c) + c) \cap (W + c)$ and $\mu(E + z) = 0$ for every $z \in Z$.²¹ The set E is shy in C if it is shy at each point $c \in C$. A (not necessarily universally measurable) subset $F \subset C$ is shy in C if it is contained in a shy universally measurable set. A subset $K \subset C$ is prevalent in C if its complement $C \setminus K$ is shy in C .*

Like Lebesgue measure 0, relative shyness and prevalence have many properties desirable for measure-theoretic notions of “smallness” and “largeness”: relative shyness is translation invariant, preserved under countable unions, and coincides with Lebesgue measure 0 in \mathbf{R}^n , and no relatively open set is relatively shy.

We note a simple but important property common to both residual and relative prevalence as notions of genericity with respect to subsets of \mathcal{C}^T .

Lemma 3 *Let $X \subset \mathcal{C}^T$ be universally measurable. If $X^c = \mathcal{C}^T \setminus X$ has a non-empty relative interior, then X is neither residual nor relatively prevalent in \mathcal{C}^T .*

Proof For relative prevalence the result is immediate from the definitions and the fact that no relatively open set is relatively shy. To see that X is not residual in \mathcal{C}^T , note that \mathcal{C}^T is a compact metric space, hence a Baire space. The conclusion then follows immediately from the Baire Category Theorem. ■

From this simple observation, we conclude that optimal full extraction is neither generically possible nor generically impossible.

²⁰Well-known problems with interpreting topological notions of genericity are illustrated by simple examples of open and dense sets in \mathbf{R}^n having arbitrarily small Lebesgue measure, and residual sets of Lebesgue measure 0.

²¹A set $E \subset Y$ is universally measurable if for every Borel measure η on Y , E belongs to the completion with respect to η of the sigma algebra of Borel sets.

Theorem 5 *Let $|T| \geq |S|$. Neither \mathcal{O} nor \mathcal{N} is residual in \mathcal{C}^T . Neither \mathcal{O} nor \mathcal{N} is relatively prevalent in \mathcal{C}^T .*

Proof Both \mathcal{O} and \mathcal{N} are Borel sets, hence are universally measurable. By definition, $\mathcal{O} \cap \mathcal{N} = \emptyset$, so $\mathcal{O} \subset \mathcal{N}^c$ and $\mathcal{N} \subset \mathcal{O}^c$. The results will all follow provided both \mathcal{O} and \mathcal{N} have non-empty interior. In both cases, we will establish this by constructing interior points.

First consider \mathcal{O} . Choose $\{\pi(t) \in \Delta : t \in T\}$ such that $\pi(t) \notin \text{co} \{\pi(t') : t' \neq t\}$ for each t , i.e., such that $\{\{\pi(t) : t \in T\}$ satisfies (OFE). Choose $\epsilon > 0$ such that

$$\forall t \in T : B_\epsilon(\pi(t)) \cap \text{co} \left\{ \bigcup_{t' \neq t} B_\epsilon(\pi(t')) \right\} = \emptyset$$

that is, such that $\{B_\epsilon(\pi(t)) : t \in T\}$ also satisfies (OFE). Now if $\Pi \in \mathcal{C}$ and $d(\Pi, \{\pi(t)\}) < \epsilon$, $\Pi \subset B_\epsilon(\pi(t))$.²² Thus any collection $\{\Pi(t) \in \mathcal{C} : t \in T\}$ such that $d(\Pi(t), \{\pi(t)\}) < \epsilon$ for each t must satisfy (OFE) as well. From this we conclude that $\{\pi(t) \in \Delta : t \in T\}$ is an interior point of \mathcal{O} .

Next, consider \mathcal{N} . Fix $t_0 \in T$. Choose $\Pi(t_0) \in \mathcal{C}$ such that $\text{rint}\Pi(t_0) \neq \emptyset$. Fix $\epsilon > 0$ and choose $\bar{\pi}(t_0) \in \text{rint}\Pi(t_0)$ such that $B_\epsilon(\bar{\pi}(t_0)) \subset \Pi(t_0)$. Now choose $\{\Pi(t) \in \Delta^c(S) : t \in T \setminus \{t_0\}\}$ such that

$$\text{co} \left\{ \bigcup_{t \neq t_0} \Pi(t) \right\} \subset B_{\epsilon/4}(\bar{\pi}(t_0))$$

In particular then, $\{\Pi(t) : t \in T\}$ violates (NOFE), so $\{\Pi(t) : t \in T\} \in \mathcal{N}$.

If $\{\tilde{\Pi}(t) \in \mathcal{C} : t \in T \setminus \{t_0\}\}$ is any collection such that $d(\tilde{\Pi}(t), \Pi(t)) < \epsilon/4T$ for each $t \neq t_0$, then

$$\text{co} \left\{ \bigcup_{t \neq t_0} \tilde{\Pi}(t) \right\} \subset B_{\epsilon/2}(\bar{\pi}(t_0))$$

Finally, if $\Pi \in \mathcal{C}$ and $d(\Pi, \Pi(t_0)) < \epsilon/2$, then $B_{\epsilon/2}(\bar{\pi}(t_0)) \subset \Pi$. Putting these observations together, any collection $\{\tilde{\Pi}(t) \in \mathcal{C} : t \in T\}$ such that $d(\tilde{\Pi}(t_0), \Pi(t_0)) < \epsilon/2$ and $d(\tilde{\Pi}(t), \Pi(t)) < \epsilon/4T$ for each $t \neq t_0$ will also violate (NOFE) and hence will belong to \mathcal{N} . ■

Taken together, Theorems 4 and 5 show that the conditions for optimal full extraction are more stringent in a world with Knightian uncertainty than the generic conditions for full rent extraction in Bayesian mechanisms. As argued above, optimality makes incentive constraints harder to satisfy when Knightian uncertainty is introduced, so it is natural that full extraction would correspondingly become more difficult. These results also suggest that concerns about the possibility of full extraction and the existence of information rents translate naturally to a choice between methods of resolving incentive compatibility with Knightian uncertainty.

²²Here $d(A, B)$ denotes the Hausdorff distance between $A, B \in \mathcal{C}$, defined by

$$d(A, B) = \max \left\{ \sup_{x \in A} \text{dist}(x, B), \sup_{y \in B} \text{dist}(y, A) \right\}.$$

6 Conclusion

We have developed a framework for mechanism design in the presence of Knightian uncertainty. In this setting, the distinction between maximal and optimal incentive compatibility is crucial. Under maximal incentive compatibility, the introduction of Knightian uncertainty weakens incentive constraints, resulting in predictions that closely resemble standard Bayesian mechanism design. Things are very different under optimal incentive compatibility, as the introduction of Knightian uncertainty imposes more stringent incentive constraints. Without uncertainty, the mechanisms that satisfy interim and ex-post incentive compatibility are significantly different. In contrast, the introduction of a (possibly small amount) of uncertainty implies that only ex-post incentive compatible mechanisms may be robust to small perturbations in standard models.

Similarly, with Knightian uncertainty, full extraction of information rents in optimal incentive compatible mechanisms requires sufficient heterogeneity of beliefs across types, while in the absence of Knightian uncertainty, full extraction of these rents is possible even when beliefs are arbitrarily close.

These results contrast with recent work in the robust mechanism design literature while reaching qualitatively similar conclusions. We maintain standard assumptions regarding common knowledge for all participants, while allowing for Knightian uncertainty. In contrast, following Wilson (1987), much recent work has taken up the charge of the oft-quoted “Wilson doctrine” to consider robustness to weakening common knowledge assumptions. Our results indicate that even maintaining standard assumptions regarding common knowledge, simple ex-post mechanisms may emerge as the only feasible mechanisms robust to Knightian uncertainty.

Appendix: Direct Mechanisms and the Revelation Principle

In this appendix, we define the general class of mechanisms and verify that the revelation principle holds for both maximal and optimal strategies so that our focus on direct mechanisms is justified.

For simplicity we focus on the single agent case. The notation and basic setup are as in Section 3. The outcome space is O , the state space is S , the type space is T , and the agent has type-dependent ex post payoff

$$u : O \times T \times S \rightarrow \mathbf{R}.$$

A mechanism in this environment consists of a message space B and a function $g : B \times S \rightarrow O$. If the agent chooses to participate in the mechanism, she chooses a message $b \in B$, and the mechanism specifies a state-dependent outcome $g(b, s)$. Moral hazard is ruled out: the agent always obeys the mechanism. A mixed strategy is a function $\sigma : T \rightarrow \Delta(B)$. For each type $t \in T$ there is a closed, convex set of beliefs $\Pi(t) \subset \Delta(S)$ such that

$$\sigma(t) \succeq_t \sigma'(t) \text{ if and only if } E_\pi [E_{\sigma(t)} [u(g(b, s), t, s)]] \geq E_\pi [E_{\sigma'(t)} [u(g(b, s), t, s)]] \quad \text{for all } \pi \in \Pi(t).$$

We consider two notions of best response: a maximal strategy corresponds to the requirement that no alternative does better, while an optimal strategy must be better than any other feasible strategy.²³

Definition 13 *A strategy σ is maximal if there is no other strategy $\sigma' : T \rightarrow \Delta(B)$ such that*

$$E_\pi [E_{\sigma'(t)} [u(g(b, s), t, s)]] > E_\pi [E_{\sigma(t)} [u(g(b, s), t, s)]] \quad \text{for all } \pi \in \Pi(t).$$

Definition 14 *A strategy σ is optimal if for each $\sigma' : T \rightarrow \Delta(B)$:*

$$E_\pi [E_{\sigma(t)} [u(g(b, s), t, s)]] \geq E_\pi [E_{\sigma'(t)} [u(g(b, s), t, s)]] \quad \text{for all } \pi \in \Pi(t).$$

A social choice function $\psi : T \times S \rightarrow O$ specifies a feasible outcome for any pair (t, s) .

Definition 15 *A mechanism g implements the social choice function ψ in maximal (optimal) strategies if there exists a maximal (optimal) pure strategy $\beta : T \rightarrow B$ such that $g(\beta(t), s) = \psi(t, s)$ for all $t \in T$ and for all $s \in S$.*

Our analysis will focus on truth-telling in direct mechanisms. At truth-telling, the ex-post utility of the agent is:

$$u(g(t, s), t, s)$$

With this in mind, we can define mechanisms that implement truth-telling as follows.

²³The notion of maximal best response in games with incomplete preferences, and the corresponding notion of Nash equilibrium, was introduced by Shapley (1959) and Aumann (1962).

Definition 16 *A social choice function ψ is truthfully implementable in maximal strategies if there exists a mechanism g such that truth-telling is a maximal strategy in g and $g(t, s) = \psi(t, s)$ for each $t \in T$ and all $s \in S$.*

Equivalently, ψ is truthfully implementable in maximal strategies if for each $t \in T$ there exists no $\sigma'(t) \in \Delta(B)$ such that

$$E_\pi [E_{\sigma'(t)} [u(g(b, s), t, s)]] > E_\pi [u(g(t, s), t, s)] \quad \text{for all } \pi \in \Pi(t)$$

Definition 17 *A social choice function ψ is truthfully implementable in optimal strategies if there exists a mechanism g such that truth-telling is an optimal strategy in g and $g(t, s) = \psi(t, s)$ for each $t \in T$ and all $s \in S$.*

Thus if ψ is truthfully implementable in optimal strategies, then for each $t \in T$ and for all $\sigma'(t) \in \Delta(B)$

$$E_\pi [u(g(t, s), t, s)] \geq E_\pi [E_{\sigma'(t)} [u(g(b, s), t, s)]] \quad \text{for all } \pi \in \Pi(t).$$

With these formalities in place, a version of the revelation principle follows.

Proposition 1 (The Revelation Principle) *If a social choice function ψ can be implemented in maximal (optimal) strategies by a mechanism g , then ψ is also truthfully implementable in maximal (optimal) strategies.*

Proof We prove the result for the case of maximal strategies; for optimal strategies the argument is analogous. By assumption, there exists a maximal pure strategy β such that $g(\beta(t), s) = \psi(t, s)$. In particular, for each $t \in T$ there is no $\sigma'(t) \in \Delta(B)$ such that

$$E_\pi [E_{\sigma'(t)} [u(g(b, s), t, s)]] > E_\pi [u(g(\beta(t), s), t, s)] \quad \text{for all } \pi \in \Pi(t).$$

But $u(g(\beta(t), s), t, s) = u(\psi(t, s), t)$, which implies by definition that ψ is truthfully implementable. ■

References

- AHN, D. (2007): "Hierarchies of Ambiguous Beliefs," *Journal of Economic Theory*, 136.
- AKERLOF, G. A. (1970): "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, 84, 488–500.
- ANDERSON, R. M., AND W. R. ZAME (2001): "Genericity with Infinitely Many Parameters," *Advances in Theoretical Economics*, 1, Article 1.
- AUMANN, R. J. (1962): "Utility Theory without the Completeness Axiom," *Econometrica*, 30, 445–462.
- (1964): "Utility Theory without the Completeness Axiom: A Correction," *Econometrica*, 32, 210–212.
- BERGEMANN, D., AND S. MORRIS (2005): "Robust Mechanism Design," *Econometrica*, 73, 1771–1813.
- BEWLEY, T. F. (1986): "Knightian Decision Theory: Part I," Discussion paper, Cowles Foundation.
- (2002): "Knightian Decision Theory: Part I," *Decisions in Economics and Finance*, 2, 79–110.
- BIKHCHANDANI, S., S. CHATTERJI, R. LAVI, A. MU'ALEM, N. NISAN, AND A. SEN (2006): "Weak Monotonicity Characterizes Deterministic Dominant-Strategy Implementation," *Econometrica*, 74, 1109–1132.
- BOSE, S., E. OZDENOREN, AND A. PAPE (2006): "Optimal Auctions with Ambiguity," *Theoretical Economics*, 1, 411–438.
- CHEN, Y., P. KATUSCAK, AND E. OZDENOREN (2007): "Sealed Bid Auctions with Ambiguity: Theory and Experiments," *Journal of Economic Theory*, 136, 513–535.
- CHRISTENSEN, J. P. R. (1974): *Topology and Borel Structure*. Amsterdam: North Holland.
- CHUNG, K.-S., AND J. ELY (2007): "Foundations of Dominant Strategy Mechanisms," *Review of Economic Studies*, 74, 447–476.
- CRÉMER, J.-J., AND R. MCLEAN (1985): "Optimal Selling Strategies under Uncertainty for a Discriminatory Monopolist when Demands Are Interdependent," *Econometrica*, 53, 345–61.
- (1988): "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, 56, 1247–57.
- DUBRA, J., F. MACCHERONI, AND E. A. OK (2004): "Expected Utility Theory without the Completeness Axiom," *Journal of Economic Theory*, 115, 118–133.

- ELLSBERG, D. (1961): "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics*, 75, 643–669.
- GHIRARDATO, P., F. MACCHERONI, AND M. MARINACCI (2004): "Differentiating Ambiguity and Ambiguity Attitude," *Journal of Economic Theory*, 118, 133–173.
- GHIRARDATO, P., F. MACCHERONI, M. MARINACCI, AND M. SINISCALCHI (2003): "A Subjective Spin on Roulette Wheels," *Econometrica*, 71, 1897–1908.
- GILBOA, I., F. MACCHERONI, M. MARINACCI, AND D. SCHMEIDLER (2008): "Objective and Subjective Rationality in a Multiple Prior Model," Discussion paper.
- GILBOA, I., AND D. SCHMEIDLER (1989): "Maxmin Expected Utility with Non-unique Prior," *Journal of Mathematical Economics*, 18, 141–153.
- GIROTTO, B., AND S. HOLZER (2005): "Representation of Subjective Preferences Under Ambiguity," *Journal of Mathematical Psychology*, 49, 372–382.
- HARSANYI, J. C. (1967): "Games with Incomplete Information Played by 'Bayesian' Players," *Management Science*, 14, 159–182 320–334 486–502.
- HEIFETZ, A., AND Z. NEEMAN (2006): "On the Generic Impossibility of Full Surplus Extraction in Mechanism Design," *Econometrica*, 74, 213–233.
- HUNT, B., T. SAUER, AND J. YORKE (1992): "Prevalence: A Translation Invariant 'Almost Every' on Infinite Dimensional Spaces," *Bulletin (New Series) of the American Mathematical Society*, 27, 217–238.
- KNIGHT, F. H. (1921): *Uncertainty and Profit*. Boston: Houghton Mifflin.
- KRISHNA, V. (2002): *Auction Theory*. Academic Press.
- LEDYARD, J. O. (1978): "Incentive compatibility and Incomplete Information," *Journal of Economic Theory*, 18, 171–189.
- (1979): "Dominant Strategy Mechanisms and Incomplete Information," in *Aggregation and Revelation of Preferences*, ed. by J.-J. Laffont. Amsterdam: North-Holland, chap. 1979.
- LEVIN, D., AND E. OZDENOREN (2004): "Auctions with uncertain numbers of bidders," *Journal of Economic Theory*, 118, 229–251.
- MCAFEE, P. R., AND P. J. RENY (1992): "Correlated Information and Mechanism Design," *Econometrica*, 60, 395–421.
- MERTENS, J.-F., AND S. ZAMIR (1985): "Formalization of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 1214, 1–29.
- MILGROM, P., AND I. SEGAL (2002): "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70, 583–601.

- NEEMAN, Z. (2004): “The Relevance of Private Information in Mechanism Design,” *Journal of Economic Theory*, 117, 55–77.
- OK, E. A. (2002): “Utility Representation of an Incomplete Preference Relation,” *Journal of Economic Theory*, 104, 429–449.
- RIGOTTI, L., AND C. SHANNON (2005): “Uncertainty and Risk in Financial Markets,” *Econometrica*, 73, 203–243.
- RIGOTTI, L., C. SHANNON, AND T. STRZALECKI (2008): “Subjective Beliefs and Ex-Ante Trade,” *Econometrica*, 76(5), 1167–1190.
- ROCHET, J.-C. (1987): “A necessary and sufficient condition for rationalizability in a quasi-linear context,” *Journal of Mathematical Economics*, 16, 191–200.
- SCHMEIDLER, D. (1989): “Subjective Probability and Expected Utility without Additivity,” *Econometrica*, 57(3), 571–587.
- SHAPLEY, L. S. (1959): “Equilibrium Points in Games with Vector Payoffs,” *Naval Research Logistics Quarterly*, 6, 57–61.
- SHAPLEY, L. S., AND M. BAUCCELLS (2008): “Multiperson Utility,” *Games and Economic Behavior*, 62, 329–347.
- WILSON, R. (1987): “Game-Theoretic Analyses of Trading Processes,” in *Advances in Economic Theory: Fifth World Congress*, ed. by T. Bewley. Cambridge: Cambridge University Press.