Robustness in Mechanism Design and Contracting

Gabriel Carroll

Stanford University; email: gdc@stanford.edu

Version: November 5, 2018

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–31

https://doi.org/10.1146/(doi to be added)

Keywords

Contract theory, mechanism design, non-Bayesian modeling, robustness, uncertainty, worst-case

Abstract

This survey summarizes a nascent body of theoretical research on design of incentives when the environment is not fully known to the designer, and offers some general lessons from the work so far. These newer models based on uncertainty and robustness offer an additional set of tools in the toolkit, complementary to more traditional, fully-Bayesian modeling approaches, and broaden the range of problems that can be studied. The kinds of insights that such models can offer, and the methodological and technical challenges they confront, broadly parallel those of traditional approaches.

1. INTRODUCTION

This survey gives an overview of recent theory on robust design of incentives when the designer does not know all the details of the environment.

Traditional models also usually assume that the designer is not fully informed, and agents have some private information. But the traditional approach in economics is to assume that all uncertainty can be described by a probabilistic belief. What distinguishes the work surveyed here is that at least some of the uncertainty is non-probabilistic. Thus the designer must evaluate possible mechanisms by some non-Bayesian criterion. Such models have sometimes been viewed as exotic, but a general theme that will emerge is that these newer models can serve the same range of purposes as traditional Bayesian ones — such as understanding why certain incentive mechanisms are common, designing new ones, or understanding the intrinsic limits posed by incentive considerations — and can sometimes do so in cleaner and more intuitive ways.

The survey is intended both to give a succinct guide to the somewhat dispersed work that currently exists in this area, and to try to draw some general lessons from the efforts so far. Accordingly, it is intended not only for researchers specifically looking to contribute to these efforts, but also more broadly for anyone interested in current conceptual tools for thinking about incentives — economic theorists, as well as scholars in adjoining areas where incentive design is important, such as industrial organization, corporate finance, political economy, and theoretical computer science. The survey will not assume specific technical background in mechanism design and contract theory.

A couple quick notes on terminology: First, there does not seem to be universal agreement on how "mechanism design" and "contract theory" are delineated, or the extent of the overlap. Here, no particular distinction will be made. The label "mechanism design" will be applied broadly, to refer to the study of designed interactions with a focus on the strategic incentives they create. The word "contract" will be used in some applications, but for reasons of convention rather than principle. Second, it bears mention that the specific phrase "robust mechanism design" has been used by some authors with slightly different, though overlapping, meanings to that used here (e.g. Börgers 2015, Chapter 10).

1.1. Non-Robust Models

To illustrate some of the motivation for robust modeling in mechanism design, it will help to begin with a few examples of *non*-robust mechanisms at the core of the traditional canon.

1.1.1. Moral Hazard. In the classic formulation of a moral hazard model (e.g. Holmström 1979), a principal hires an agent to exert effort which then produces a stochastic amount of output for the principal. In particular, the agent is to choose an action a, which consists of either exerting high effort or low effort, $a \in \{H, L\}$.[1] Output y then follows a distribution that depends on the effort level, $F(y|a)$, with density $f(y|a)$. The principal cannot observe effort directly, but can observe output, and can write a contract $w(y)$, specifying remuneration to the agent as a function of output. The principal's payoff is $g(y - w)$; the agent's is $u(w) - c(a)$. Here $g(\cdot)$ and $u(\cdot)$ are utility functions, and $c(\cdot)$ is a cost-of-effort function, with $c(H) > c(L)$. The principal's problem is to choose the function $w(y)$ opti-

[1] The original formulation in Holmström allowed a to be a continuous one-dimensional choice. The ideas are similar, but the mechanics are more transparent in the binary-effort case.

mally, taking into consideration that the agent's optimal choice of action a responds to the incentives provided by the contract $w(y)$. Assuming that the parameters are such that the optimal contract induces high effort $a = H$, the contract is characterized as maximizing $\int g(y - w(y))\, dF(y|H)$, subject to two constraints: an incentive constraint, that the agent indeed prefers to exert effort H rather than L; and a participation constraint, that the agent's expected payoff cannot fall below some exogenous value $\underline{u}$, representing his outside option.

The solution to this problem satisfies the relation

$$\frac{g'(y - w(y))}{u'(w(y))} = \lambda + \mu \cdot \left(1 - \frac{f(y|L)}{f(y|H)}\right) \qquad \text{for all } y, \qquad 1.1.$$

where λ and μ are (endogenously determined) Lagrange multipliers on the participation constraint and the incentive constraint, respectively. Under standard assumptions (such as risk-aversion), once λ and μ are pinned down, this fully characterizes the contract.

Of particular importance, the fraction on the right-hand side is a likelihood ratio; it captures how informative y is as a signal that the agent was exerting the intended level of effort. (Note that this is an interpretation of the algebra; in equilibrium, the agent would always choose high effort.) The equation (1.1) shows that, all else equal, it is optimal to provide incentives by paying more for realizations of output that are a stronger signal of the agent having taken the target action H.

Essentially, we optimally provide incentives by rewarding the agent based on whether it looks like he has done the right thing, regardless of whether the realized outcome was a good one. This insight has been fundamental to the development of contract theory.

Yet, in reality, when we see explicit pay-for-performance contracts, they do not have likelihood ratios in them. Conversely, we see a few common forms of contracts — such as **linear** contracts, or **bonus** contracts — used across a wide range of situations, where there is no particular reason to expect likelihood ratios to be similar. Moreover, if we wished to write a contract based explicitly on formula (1.1), we would be hard-pressed to do so, given not only the strained assumption of only two possible effort choices (or even one-dimensional effort as in Holmström's formulation), but also the requirement that the designer be able to precisely specify the densities $f(y|H)$ and $f(y|L)$, as well as the utility functions. All this suggests that if we wish to explain how real-world incentive contracts are written, or give detailed advice on how they should be written, we should adopt a modeling framework that reflects these limits on the plausible knowledge of the designer. Even if we are not committed to a literal interpretation of the results, we might consider such a framework and see whether it can deliver new insights.

linear contract: contract that pays the agent a constant fraction of output

bonus contract: contract that gives a discrete payment if output exceeds some fixed threshold

1.1.2. Auctions with Correlated Values.

Our second example comes from auction theory. Consider a seller, with an object available to sell, who would like to make as much money as possible (in expectation). There are two risk-neutral buyers. Each has a value for the object, drawn independently from some distribution, and each buyer privately knows her own value. For specificity, let's assume each value is drawn uniformly between 0 and 1.

The first-best from the seller's point of view would be to find out each buyer's willingness to pay for the object, choose the buyer with higher value, and sell it to her at a price equal to her value. But the seller cannot achieve this by simply asking the buyers their values, since each buyer would lie to get a better price. More generally, no matter what mechanism the seller proposes, her ability to extract revenue is limited by incentive constraints, resulting from each buyer's ability to strategically behave as if her value is lower than it actually is.

The classic analysis of Myerson (1981) shows how to formalize the seller's problem and derive the revenue-maximizing auction. There are actually many ways to write the auction rules that all lead to the same equilibrium outcome; we conventionally express this by saying that "the" optimal auction can be implemented in many ways. One such implementation is a second-price auction with a reserve price of $1/2$: that is, each buyer makes a bid; the higher bidder, if she bids above the reserve, receives the object, and is charged a price equal to the maximum of the reserve and her opponent's bid.

So far so good. Now consider a variant model: the buyers' values are no longer drawn independently. Instead, with probability $1/2$, the two buyers' values are drawn independently uniformly on $[0, 1]$ as before; with remaining probability $1/2$, just one value v is drawn from uniform $[0, 1]$, and both buyers' value equals v. Each buyer knows only her own value, and does not know which of the two cases arose.

In this model, the designer can now extract the full first-best. Here is one way: Ask each buyer to report her value. The higher bidder is sold the object, at a price equal to her reported value. (A tie — which in this example would occur with probability $1/2$ — would be broken by a coin flip.) In addition, each bidder, in order to participate in the auction, is required to accept a "side bet" in which she pays 1 to the seller if the two bidders' reports differ, but receives 1 from the seller if they are identical. In this mechanism, each bidder is willing to participate (and report truthfully): if she wins the object, she pays her value, for a profit of zero; and in the side bet, she wins and loses 1 with equal probability, which washes out given risk-neutrality. For the same reason, the seller does indeed extract the full first-best surplus. Moreover, a bidder cannot benefit by misreporting her value, because if she reports anything other than the truth, her opponent's bid has a 0% chance of being identical to her own, and so she loses the bet with probability 1. Even though she may gain from buying the object at a better price, losing the bet swamps this gain.

The possibility of using such side bets to extract the full surplus was noted already by Myerson (1981), but is usually credited to Crémer & McLean (1988) (see also McAfee & Reny 1992). It is fair to say that the auction described above is not one we would see in practice. The model makes extremely strong demands, not only on the seller's knowledge of the distribution of the values, but also on her confidence that the buyers share this knowledge (and that they have no additional information). It also leans heavily on the assumption of risk-neutrality and, for that matter, on expected-utility maximization (a standard assumption in economists' models, but far less universal in practice).

To be clear, none of the papers cited above proposed their full-surplus-extraction results as a serious prescription for practical use; Crémer & McLean (1988) and McAfee & Reny (1992) used them to provide commentaries on modeling methodology. Indeed, arguably the lasting value of these results has been to serve as a guide to modelers, showing that we need to impose some assumptions (either assume independent types, or restrict the allowable mechanisms somehow), otherwise things quickly go off the rails.

Note that both here and in the preceding contract example, the mechanism identified as optimal by the classical theory was finely tailored to detailed parametric assumptions on the environment. One driving goal in much of the robustness literature has been to see whether incorporating uncertainty on the designer's part leads to simpler or less detail-sensitive mechanisms. We shall revisit these connections in the concluding discussion.

1.1.3. Subgame-Perfect Implementation. The preceding two examples studied situations where the designer maximizes a numerical objective, such as expected revenue. A separate branch of mechanism design studies **implementation** questions, in which a designer has some target outcome (or a set of acceptable outcomes) for each situation that may arise, and seeks a mechanism that will always ensure such an outcome.

implementation problem: problem where a designer seeks a mechanism that will always ensure an acceptable outcome

The implementation problem can be formulated in many flavors. Probably the most classic version is the one generally credited to Maskin (1999) (though with antecedents such as Hurwicz 1972). There are several agents, whose preferences depend on a state of nature. The agents all know what state is realized, but the designer does not. Agents' behavior in any possible mechanism is described by Nash equilibrium. The designer has in mind a **social choice correspondence** (SCC) identifying acceptable outcomes, and wishes to design a mechanism for which Nash equilibrium play always results in such an outcome. One general-purpose solution is to ask everyone to report the state, enact an acceptable outcome if the reports all agree, and punish all the agents if they disagree. Truthful reporting is then an equilibrium. But there are also lots of other equilibria in which the agents coordinate on some false report. So Maskin (1999), and subsequent literature, takes up a more demanding goal: to design mechanisms in which every equilibrium, not just one, produces acceptable outcomes.

social choice correspondence (SCC): a specification of one or more acceptable outcomes in each state

For an example, imagine a buyer and seller who contract to trade some good that has not yet been produced. After it is produced, it may turn out to be either low quality or high quality. Assume that both parties observe the quality, and if it is low, the good is worth 40 to the buyer; if high, it is worth 60 to the buyer (and the cost to the seller is 0 in either case). Assume the parties would like to trade at a price of 20 if low quality and 30 if high, thus splitting the gains from trade. They cannot simply write a contract saying this, because such a contract cannot be enforced in court; even though both agents know the true quality, the court lacks the expertise to verify it. They could use the general-purpose mechanism above, where the buyer and seller both report the quality to a neutral third party and trade at the corresponding price if their reports agree (and don't trade if they disagree); but this is not entirely satisfactory, because the seller (say) may be worried that she cannot escape from the bad equilibrium wherein both parties just always report low quality regardless of the true state. In fact, *no* static mechanism for this problem is free from such bad equilibria: the SCC fails a key necessary property known as "Maskin monotonicity."

Moore & Repullo (1988) proposed a solution, applicable for this example and in general:[2] allow dynamic mechanisms, and assume that agents play a subgame-perfect equilibrium. They show that this allows almost any outcome to be implemented. In the above example, their construction could be carried out as follows: The buyer first makes a report of the quality. The seller then can agree, and they trade at the corresponding price; or can challenge the report. If the seller challenges a low-quality report, then the buyer is charged a large fine, and then is given the chance to buy at the higher price of 55. Note that allowing challenges indeed gets rid of the bad equilibrium. If the true quality is high, the buyer doesn't want to try to get a cheap price by reporting low quality, because the seller would challenge in order to (successfully) sell at the higher price.

However, as noted by Aghion, Fudenberg, Holden, Kunimoto & Tercieux (2012), this kind of mechanism is not robust, in that it depends very sensitively on the assumption of

[2] If the chronology seems odd, note that Maskin's paper circulated unpublished since 1977.

complete information.[3] Suppose instead that each player has just a small ϵ probability of misperceiving the quality. Suppose that the buyer still is expected to report (his perception of) the quality truthfully. If the seller thinks the quality is high, but he sees the buyer report low, then he must conclude that someone misperceived the quality — and it's not clear who. This makes him much less inclined to challenge a low-quality report, which in turn destroys the buyer's incentive to report truthfully in the first place.

1.2. Perspectives on Mechanism Design

As the examples above demonstrate, seemingly non-robust models can be quite instructive for some purposes but unreasonable for others. As we think about the possible contributions of robust models, it will help to keep in mind why we might be interested in mechanism design in the first place.

1. At the most direct level, mechanism design is truly about design — that is, it aims to give guidance to people designing allocation mechanisms or incentive systems in the real world.

   This can take place in multiple ways. In some cases, one formulates models that map quite literally onto an application being studied, and designs the rules of the mechanism at a detailed level, adapting them to specific features of the application. Much recent work in matching theory has this flavor; one classic example is Roth & Peranson (1999). So does much of the mechanism design work in algorithmic game theory (see Nisan et al. 2007).

   A different perspective is to use models simply as stylized representations, meant to deliver qualitative insights. Thus, for example, the moral hazard model above delivers a lesson — that we should reward outcomes that are indicative of doing the right thing, not good outcomes per se — even if we would not seriously contemplate a literal use of the formula that comes out of it.
2. Another view on work in mechanism design is that it provides explanations for mechanisms (or features thereof) seen in the real world, rationalizing them as optimal in various environments. Aside from the explanatory value, this perspective is also indirectly useful for design, insofar as designers often have a choice of what model to write down, and one way to evaluate a model is to see whether its predictions in already-understood situations line up with observed reality.

   This "explaining observations" view of mechanism design can again be applied either at the qualitative level, or at a more literal level, explaining specific forms of incentives. An example of the latter is Holmström & Milgrom's (1987) model to explain the pervasiveness of linear incentive contracts.
3. A very different view of mechanism design, dating back to the field's origins in Hurwicz (1972) (and re-emphasized for instance in Bergemann & Morris 2017), is metaphorical: there is no actual designer, but one studies the design problem to learn about the limits of what any actual mechanism or institution can achieve. For example, a first-order lesson from the classic Myerson-Satterthwaite (1983) bilateral trade model

[3] The distinction between "traditional" and "robust" design here fits only clumsily with the Bayesian / non-Bayesian uncertainty delineated above. But as we shall see, robustness questions studied in "numerical" mechanism design have sometimes had conceptual parallels in implementation problems, so it will be useful to discuss both.

is that, when both parties to a transaction hold private information about their own preferences, there is generally no trading mechanism that can ensure efficient outcomes. Moreover, their analysis gives a quantitative bound on the amount of inefficiency that is inevitable. This can then be useful as a benchmark to evaluate the performance of actual institutions (e.g. Larsen 2018).

4. Finally, mechanism design can provide simple modeling tools to use in the course of studying other economic phenomena. For example, the literature on organizational form surveyed in Mookherjee (2006) emphasizes agency frictions due to asymmetric information as a source of inefficiency. To compare different organizational structures, one needs to know what will happen within each structure — where the inefficiencies arise and how severe they are. Theory of optimal contracts gives a tool to write down such models and make predictions systematically, even though studying the contracts themselves is not the end goal.

The work surveyed below, on robust mechanism design for uncertain environments, can potentially contribute to each of these purposes. We will have occasion to refer back to this list periodically.

1.3. Organization

The next section dives into the robustness literature in more detail. The aim will be to represent the range of questions and models that have been studied in this literature. Sometimes this will mean that the presentation emphasizes breadth of coverage, at the expense of expositional unity. Afterwards, Section 3 will tie things back together with some general reflections.

There is much research that could fall under the heading of "robustness" but is not discussed here for space reasons. For example, a considerable body of work studies mechanisms that achieve approximate optimality (to within, say, some constant factor) across many environments. Such studies can deliver important insights, just as exact-optimality results can. This literature will not be covered here; the piece by Roughgarden and Talgam-Cohen (2019) in this issue provides an entry point for the interested reader (see also Hartline 2012). Also not covered here is work on "reduced-form" approaches to ensure robustness, such as using dominant-strategy mechanisms; see Subsection 2.2 for discussion and references.

Finally, just as game theory comes logically prior to mechanism design, so the question of making robust predictions in specific mechanisms naturally precedes the question of robust design. Yet the historical development of ideas seems to have flowed equally much in the reverse direction. But in principle one can pick any of the various dimensions of robustness considered below, and ask for robust game-theoretic analyses of "standard" mechanisms along that dimension. Some such studies have been done, but there will not be space here for a systematic discussion. For a few important works, see Neeman (2003), Battigalli & Siniscalchi (2003), Bergemann, Brooks & Morris (2017).

2. ROBUSTNESS IN MECHANISM DESIGN

The presentation of literature below will be organized loosely based on the different dimensions of the environment along which robustness may be desired. This means that there will be some hopping back and forth among different kinds of applications; for example, between work that maximizes a numerical objective and other work studying implementa-

private values: each agent knows her own preferences exactly

interdependent values: an agent's preferences can depend on information held by other agents

tion questions (as discussed above). Similarly, the bulk of the literature considers settings of **private values**, but some work considers **interdependent values** as well; both will be discussed here as is convenient.

2.1. Robustness to Technology or Preferences

To illustrate the robust approach, let us begin with the model of Carroll (2015), which studies a variant of the moral hazard problem from Subsection 1.1.1. Recall the notation from that subsection. To abstract from issues of risk-aversion, assume both parties are risk-neutral, $g(x) = u(x) = x$, and assume there is a limited liability constraint: the principal needs to write a contract satisfying $w(y) \geq 0$ for all y. In this case, there is no need for a separate participation constraint (although we could impose one, and the essential conclusions below would be unchanged). With these changes, equation (1.1) no longer applies since the optimal contract is generally a corner solution. Instead, with just two actions $\{H, L\}$ as before, the optimum puts all payment on the output with the highest likelihood ratio, and pays zero for any other output. More generally, the contract remains sensitive to assumed distributions.

Now introduce the key change: we no longer assume that the principal knows the agent's possible actions, and the resulting probability distributions over output. Instead, the set of actions available to the agent — which we call the **technology**, denoted by $\mathcal{A}$ — is unknown to the principal. When the principal contemplates any contract $w(y)$ she could offer, she evaluates it based on the expected profit that she is guaranteed to receive, no matter what the technology is. That is, she evaluates it by the worst-case criterion

$$V_P(w) = \inf_{\mathcal{A}} \left(\mathbb{E}_{F^*(w,\mathcal{A})}[y - w(y)] \right),$$

where $F^*(w, \mathcal{A})$ denotes the distribution over output that results from the agent choosing his best action, given that his technology is $\mathcal{A}$ and he faces contract w.

Of course, without any assumptions at all on the technology, no guarantee is possible — the agent might simply not be able to do anything. Instead, Carroll (2015) assumes partial knowledge: there is some set of actions, $\mathcal{A}_0$, that the principal knows the agent can take. (An action is modeled as an ordered pair, specifying the effort cost to the agent and the distribution over output that results; thus we abstract away from describing what the agent physically does, since this is not payoff-relevant.) The class of possible technologies $\mathcal{A}$ from the principal's point of view is the class of supersets of $\mathcal{A}_0$. The key result is that the optimal guarantee $V_P(w)$ is attained by a linear contract — one of the form $w(y) = \alpha y$, for some constant α. Thus, linear contracts provide the most robust way of aligning the agent's interests with the principal's.[4]

An intuition for linearity is as follows: The agent may potentially choose to produce any distribution over output, depending on the realized technology $\mathcal{A}$. From the principal's point of view, there is only one constraint to discipline the agent's choice: a lower bound on $\mathbb{E}_F[w(y)]$, coming from the known actions. (If some known action gives the agent an expected payoff of z, then the principal can infer the agent would never choose any distribution that pays less than z on average, since such an action would definitely be suboptimal.)

[4] A number of other works, most prominently Holmström & Milgrom (1987) and Diamond (1998), have given other arguments for linear contracts using Bayesian models.

On the other hand, the principal's objective is a lower bound on $\mathbb{E}_F[y - w(y)]$. Linear contracts provide a tight link between the former expectation and the latter, without needing to know anything further about F. Nonlinear contracts may provide some link as well, but generally a less tight one, leaving room for improvement.

How does one interpret the maxmin criterion? One could take it literally, as a description of a principal's decision-making. A broader viewpoint is that it is a formalization of a robustness property of linear contracts — a way in which one can make guarantees about the principal's payoff with very little information about the environment (here described by $\mathcal{A}$). This property may help explain why linear contracts are widespread in practice, even if nobody is explicitly optimizing a maxmin objective. Note that any such explanation requires that linear contracts be special in some way, which is why it is important that the guarantee criterion picks them out as optimal.

The machinery and result of the above model can readily be carried to more complex situations. Dai & Toikka (2017) consider a problem where a principal writes robust contracts for a team of agents, where each agent $i = 1, \ldots, n$ is to be paid according to a function of total output, $w_i(y)$, and output is determined jointly by the agents' actions (which may interact in an arbitrary way). The principal knows some actions available to each agent, but other, unspecified actions may also be possible. Dai and Toikka show two main results: first, to get any guarantee at all, the principal must offer contracts such that $w_i(y)$ and $w_j(y)$ are linearly related to each other for all i, j; second, the optimal contracts are linear in total output, as in the one-agent case. Marku & Ocampo Díaz (2017) consider a common agency model, where two principals simultaneously offer contracts to an agent, whose action then produces output for each principal separately. Each principal i wishes to counteract the incentives offered by principal j, since those incentives could lead the agent to produce output for j and not for i; in equilibrium, each principal offers a linear contract that is increasing in the output she receives and *decreasing* in the output for the other principal. Carroll (2018b) considers a setting of costly information acquisition: rather than producing output directly, the agent uses his technology (again, only partially known to the principal) to get information about an underlying state of nature, which the principal can then use to make some investment decision. There is a natural way to define linear contracts here: the agent acquires information, recommends an investment, and then is paid a fraction α of the ensuing returns. The maxmin-optimal contract is actually a more complicated variant in general, but is exactly linear in some special cases.

There has been other work showing how simple contracts provide robust guarantees in principal-agent settings. Hurwicz & Shapiro (1978), seemingly the earliest contracting model in this strain, considered a model in which the agent's effort cost is quadratic, with unknown coefficient. In this case, even the first-best surplus may be arbitrarily small, so no profit guarantee is possible. Instead, they considered guarantees on the ratio of the principal's actual payoff to the payoff that she could have gotten if she knew the true environment (i.e. the agent's cost function). The best such guarantee is attained by a linear contract that pays the agent half the output. Chassang (2013, Corollary 1) also gives a result on optimality of linear contracts by a maxmin-ratio criterion over a certain class of environments.

Garrett (2014) considers a version of the classic Laffont-Tirole (1986) cost-based procurement model. In that model, a government buys a good from a supplier, and then sees the supplier's report of costs to be reimbursed. The supplier has private information about his "intrinsic" cost to produce the good; but he can also exert effort to reduce the costs below this level. The government can offer a contract that specifies payment as a function of realized cost. Laffont and Tirole's analysis showed that it is generally optimal to offer a menu of many such contracts, into which the supplier can self-select based on his intrinsic cost. In Garrett's robust version, the government has a Bayesian prior over the intrinsic cost, but has maxmin-style uncertainty about the disutility-of-effort function, and knows only some lower bound $\underline{k}$ on the net efficiency gain (i.e. cost savings minus effort disutility) that the supplier can generate. The optimal menu now consists of just two contracts, one that pays a low but fixed price (letting the supplier pocket any further cost savings from his efforts) and one that reimburses costs one-for-one.

Frankel (2014) considers a model of multiple delegated decisions. An agent faces N similarly-structured decisions; in each decision i, she receives private information θ_i and takes observable action a_i. This results in payoffs $\sum_{i=1}^{N} U_P(a_i|\theta_i)$ and $\sum_{i=1}^{N} U_A(a_i|\theta_i)$ for the principal and agent respectively. (A leading example is a teacher assigning grades in a class: θ_i is student i's actual performance; a_i is the grade assigned; the school and the teacher each have preferences about how grades correspond to performance.) The principal knows her own preference U_P, which is assumed to be supermodular (higher actions preferred for higher states), and has a prior over $(\theta_1, \ldots, \theta_N)$, but does not know anything about U_A except that it is also supermodular. Frankel shows that a maxmin-optimal mechanism simply tells the agent how many times each a_i should be chosen (so the teacher is told how many A's to give, how many B's, etc.).

Thus, these various works explore many different situations in which some simple mechanism provides intuitive guarantees using only limited information about agents' preferences or technology, and furthermore show that the simple mechanism is distinguished in this respect, as formalized using a maxmin criterion.

2.2. Robustness to Beliefs or Information

dominant-strategy (strategy-proof) mechanism: one where agents report preferences, and each agent's optimal strategy is always to report her preferences truthfully, no matter what other agents report

Probably the most widely-expressed concern for robustness in mechanism design is about robustness to agents' probabilistic beliefs about each other. Many classic analyses of design problems involving multiple agents apply the concept of Bayesian equilibrium, implicitly making strong assumptions on agents' beliefs. We saw this in the discussion of Crémer-McLean (1988) side-betting mechanisms above, but Bayesian equilibrium has also been applied in many other places and often with a less critical attitude, such as the expected externality mechanism of d'Aspremont & Gérard-Varet (1979) or the optimal bilateral trading mechanism of Myerson & Satterthwaite (1983). Such strong assumptions warrant suspicion and it is natural to ask what can be done without them.

The most common way to ensure robustness to beliefs is to design **dominant-strategy** (or **strategy-proof**) mechanisms. Such a mechanism can be analyzed without any assumptions on beliefs. A long tradition in social choice simply takes as axiomatic that mechanisms should satisfy this property — going back to the Gibbard-Satterthwaite impossibility theorem (Gibbard 1973, Satterthwaite 1975), which essentially says that if no restrictions on preferences are assumed, then the only dominant-strategy, non-randomized voting mechanisms are dictatorships; as well as more positive results in other domains (second-price auctions are one well-known example). The dominant-strategy approach is especially natural in domains such as voting or matching, where the space of preferences is unstructured and there is no obvious way to formulate a prior; and has recently gained resurgence with popular applications such as school choice, where the dominant-strategy property seems to be easier to explain to policymakers than Bayesian analyses.

The literature on what dominant-strategy mechanisms can and cannot accomplish in various settings is vast. There exist already extensive surveys, see for example Barberà (2011) or Sprumont (1995); so this work, while important, will not be detailed here. What will be addressed here, however, is some newer research on foundations for the dominant-strategy property. This work asks: the dominant-strategy property is clearly *sufficient* for robustness to beliefs, but is it *necessary*? If not, when can we do better without it? After all, in principle the designer should be able to make at least a partial prediction about what agents will do in any mechanism, dominant-strategy or not, and evaluate each mechanism based on these predictions.

To see why belief-robustness need not require dominant strategies, consider the following implementation example based on Bergemann & Morris (2005).[5] Suppose that there are two agents, who each privately know their own preference type; agent 1's type may be either θ_1 or θ_1', and likewise agent 2 may be θ_2 or θ_2'. The planner needs to choose one of six outcomes a, b, c, a', b', c'. Specifically, the planner wishes to ensure an outcome that depends on the agents' types as follows:

	θ_2	θ_2'
θ_1	a or b	a' or b'
θ_1'	c	c'

Meanwhile, the agents have the following payoffs from each outcome:

1 :

	a	b	c	a'	b'	c'
θ_1	2	−1	0	−1	2	0
θ_1'	0	0	1	0	0	1

2 :

	a	b	c	a'	b'	c'
θ_2	1	3	1	0	2	0
θ_2'	2	0	0	3	1	1

The planner cannot ensure a desirable outcome using a dominant-strategy mechanism: to make sure agent θ_1 never wants to misreport type θ_1', she would have to specify outcome a (not b) at profile (θ_1, θ_2), and b' (not a') at (θ_1, θ_2'); but then agent θ_2 would want to misreport as θ_2' when she expects 1 to report θ_1.

However, the planner can still ensure a good outcome regardless of beliefs, using the following non-dominant-strategy mechanism: agent 1 chooses one of the pairs $\{a, a'\}, \{b, b'\}$ or $\{c, c'\}$; then agent 2 chooses an outcome from 1's pair. This works because in any pair, type θ_2 would be willing to choose the unprimed outcome and θ_2' the primed outcome; and agent 1, foreseeing this, chooses

- pair $\{a, a'\}$ if she has type θ_1 and believes 2 is more likely θ_2 than θ_2';
- pair $\{b, b'\}$ if she has type θ_1 and believes θ_2 less likely than θ_2';
- pair $\{c, c'\}$ if she has type θ_1'.

Bergemann & Morris (2005) give several such examples that further distinguish among various degrees of robustness to beliefs, and then present several versions of (fairly restrictive) sufficient conditions under which implementation for all possible beliefs and higher-order beliefs is indeed equivalent to dominant-strategy implementation.

Börgers & Smith (2014) look at a more concrete context, a voting model, and likewise argue that requiring dominant strategies is too restrictive. They consider a setting

[5] The presentation here combines features of their Example 1 and Example 2.

random dictatorship: a voter is chosen at random according to an exogenously fixed distribution, and gets to pick his favorite outcome

where voters have cardinal utilities over outcomes, and voting mechanisms may be randomized. The Gibbard-Satterthwaite impossibility theorem has a generalization to such environments, due to Hylland (1980): essentially the only dominant-strategy mechanisms are **random dictatorships**. Börgers and Smith consider an alternative mechanism, in which the voters may seek "compromise" outcomes if their preferences differ, but each voter can unilaterally veto and force a return to random dictatorship. They observe that equilibrium play of this mechanism weakly dominates pure random dictatorship, in the sense that for all preferences and beliefs a voter might have, he gets at least as high an expected utility as he would under random dictatorship (since he can always enforce random dictatorship as a fallback) and for some beliefs he does strictly better. Thus, they argue, relaxing the dominant-strategy criterion overturns Hylland's negative result.

This leaves open the question of whether the gap between belief-robustness and dominant-strategy mechanisms arises in models with a numerical objective, such as revenue maximization. Chung & Ely (2007) take up the problem of auctions with correlated values, as in Subsection 1.1.2 above, asking whether a desire for robustness to agents' beliefs would justify using a dominant-strategy auction mechanism. Thus, they consider a seller who has a (correlated) prior belief over buyers' values, but does not know the buyers' beliefs about each other, and wants to maximize worst-case expected revenue, where the worst case is over beliefs the buyers may have. Under a regularity condition on the prior, they show that indeed the seller cannot do better than the optimal dominant-strategy auction. They show this by constructing particular worst-case beliefs for each agent under which the seller cannot do better than a dominant-strategy auction.[6] They also show that the result is sensitive to the regularity assumption: without the assumption, they give an example where, for any hypothesis that the seller may entertain about the buyers' beliefs, there is a side-betting mechanism that does strictly better than the best dominant-strategy mechanism.

ex post implementation: an analogue of dominant-strategy implementation for interdependent values

The above examples considered private-values environments. When values are interdependent, dominant strategies make little sense. There is an analogue, **ex post** implementation; it requires that each agent should have a prescribed strategy, specifying his action in the mechanism as a function of his (partial) information about payoffs, that is optimal regardless of others' information as long as the others follow their prescribed strategies too. But this notion has two drawbacks. First, it effectively assumes that the designer can specify exactly the nature of agents' information about payoffs, so only their beliefs about each other's information are unspecified; it is hard to think of many practical situations (other than private-values ones) where this is realistic. Second, Jehiel, Meyer-ter-Vehn, Moldovanu & Zame (2006) have shown that in a broad class of environments, the only ex post mechanisms are trivial ones. Given the weaknesses of this approach to belief-robustness, maxmin again appears as an especially natural alternative.

Brooks & Du (2018) use this approach to study robustness to beliefs in pure common-value actions.[7] That is, they now assume the value of the good to all buyers is the same, but each buyer has only noisy information about this value. Their auctioneer wants to maximize worst-case expected revenue, where the expectation is with respect to a (fixed)

[6] Specifically, in this worst case, when agent i has value θ_i, his belief about the others' values is the same as the true distribution conditional on i's value being at least θ_i (instead of exactly θ_i). See Chen & Li (2017), Yamashita & Zhu (2017) for generalizations.

[7] This work builds on earlier work by Bergemann, Brooks & Morris (2016) and Du (2018).

prior over the good's value, and the worst case is over information structures, describing what the buyers know. Their technically deep analysis identifies an optimal mechanism in which bidders make one-dimensional bids, and the object is randomly allocated to each bidder with probability proportional to his bid (or sometimes withheld); payments are given by rather involved formulas. They also identify a worst-case information structure, which has the following form: each agent i privately observes a signal s_i; these signals are exponentially distributed, independently across agents, and the good's value is correlated with them such that it depends only on the sum $\sum_i s_i$. One possible view of this paper — in line with item 3 from our list of goals in Subsection 1.2 — is that it identifies the kinds of informational environments where revenue extraction is most difficult. Bergemann, Brooks & Morris (2018) consider this same model but restrict to a class of "standard" auction formats, and show that within this class, the first-price auction is maxmin-optimal.

Carroll (2016) considers a related problem of informational robustness under interdependent values in a simpler environment, with binary outcomes (a proposed agreement can be adopted or not) and possible adverse selection. The main focus there is on one specific mechanism, in which each player just accepts or rejects the agreement, and it is only adopted if both accept; the paper examines possible equilibrium outcomes when the analyst does not know the information structure. However, a corollary of the analysis is that this mechanism is actually maxmin-optimal (as opposed to other, more elaborate exchanges of information) for some parameters.

Returning to private-value settings, Carroll (2018a) considers a planner designing a mechanism for parties to trade, who not only is unsure about the traders' information about each other, but is also concerned about resource costs spent in getting there — either in acquiring information, or influencing the information of others. Note that dominant-strategy mechanisms give no incentive for such information manipulations of any kind. The planner in this model has a prior over agents' values, and aims to maximize worst-case expected welfare; here welfare reflects surplus in the trading mechanism as well as costs spent manipulating information beforehand, and the worst case is over the "information games" that agents may have available. In a simplified bilateral trade environment, the maxmin mechanism can be identified: for some parameters, it is a dominant-strategy mechanism; for others, it is a non-dominant-strategy mechanism where one party chooses a take-it-or-leave-it price offer to the other.

To summarize briefly: although dominant-strategy (or ex post) implementation is the traditional way to ensure belief-robustness, it may not be required, and one can sometimes do better without it. Even when one cannot, this may require substantial work to prove. There are known cases where one can do better with simple constructions, but it is not always known how much better; only in a few such models has the maxmin problem been fully solved.

2.3. Robustness to Strategic Behavior

The work described in the previous subsection takes a "structural uncertainty" approach, which follows the orthodoxy of assuming equilibrium behavior, and models the designer's uncertainty as being about the underlying primitives (in this case, agents' beliefs about each other's types). An alternative is a "strategic uncertainty" approach, which relaxes equilibrium to be more agnostic about agents' strategic behavior. For example, a designer might only be confident that agents will not play strategies that are weakly dominated;

thus she would want to ensure desirable outcomes for all such strategy profiles. One might argue that this approach does not fall under the heading of "robust mechanism design" as initially delineated in Section 1, since the uncertainty is not about the specification of the environment but rather about agents' behavior within the environment. However, this literature is thematically related to that on structural uncertainty and so it makes sense to discuss them together. Moreover, in some formulations the two approaches are actually equivalent; see Bergemann & Morris (2011) and Yamashita (2015b).

Börgers (1991) first showed that explicit modeling of strategic uncertainty can be more permissive than simply requiring dominant-strategy mechanisms. Consider a voting setting: there are several possible outcomes, and each agent (voter) may have any arbitrary preference ordering over the outcomes. Again, the Gibbard-Satterthwaite theorem says that the only dominant-strategy voting mechanisms are dictatorships. But as Börgers points out, with three possible outcomes, **approval voting** (with appropriate tie-breaking provisions) guarantees a Pareto-efficient outcome whenever the voters play undominated strategies, and treats the voters fairly, thus overcoming the pessimistic conclusion of Gibbard-Satterthwaite. (A voter's undominated strategies are either to approve his most-preferred outcome or his two most-preferred outcomes.)

approval voting: each voter may approve any subset of outcomes, and whichever outcome gets the most approvals wins

There are other works exploring both the possibilities and limits of undominated-strategy implementation. Jackson (1992) identified a necessary condition on outcomes to be implementable (barring pathological mechanisms), termed "strategy-resistance," a weakening of strategy-proofness. Babaioff, Lavi & Pavlov (2009) study certain combinatorial auction settings where no nontrivial dominant-strategy mechanism is known; they offer a mechanism that provides a nontrivial welfare guarantee in undominated strategies. Börgers & Smith (2012), parallel to Börgers & Smith (2014), give examples of settings where any dominant-strategy mechanism is weakly dominated by another mechanism, in the sense that the latter does at least as well for all type profiles and can do better for some, as long as agents play undominated strategies.

One problem that has been much studied is the implementation problem as in Subsection 1.1.3, where the designer now wants to ensure an acceptable outcome as long as agents play (some version of) undominated strategies. A focal case is that of a **social choice function**, specifying just one acceptable outcome in each state. Which such functions can be implemented? This question can be asked in many different flavors, such as assuming only one round of deletion of dominated strategies, or iterated deletion; and assuming complete or incomplete information. Known answers are often technical, but two results stand out as worth mentioning here. First, Abreu & Matsushima (1992a) consider iterated deletion and complete information, and allow randomized mechanisms. They show that in many environments, any social choice function at all can be **virtually** implemented. This seemingly too-positive result has sparked debate as to what are reasonable behavioral assumptions (see Glazer & Rosenthal 1992, Abreu & Matsushima 1992b). Second, Bergemann & Morris (2009a) show how restricting the environment may help cut through the thicket. They focus on a particular class of environments, in which each agent's preferences depend only on a one-dimensional aggregate of all agents' types. They show that if the planner's desired outcome can be implemented at all, then it can be done by the direct mechanism where agents just report their types.[8]

social choice function (SCF): an SCC specifying a unique outcome in each state

virtual implementation: implementation of a desired outcome with probability at least $1 - \epsilon$, for arbitrarily small ϵ

[8] For more on this topic, see Abreu & Matsushima (1994), Bergemann & Morris (2009b, 2011), Bergemann, Morris & Tercieux (2011). The survey by Jackson (2001) discusses many other versions

Yamashita (2015a) takes the strategic uncertainty approach to optimize a numerical objective. He considers a bilateral trade model, where the designer has a prior over the buyer's and seller's values but the only behavioral assumption is undominated strategies; thus the objective is maxmin expected welfare, where the min is over undominated strategies (and the expectation is with respect to the prior over values). For some prior distributions, the designer can do no better than the optimal dominant-strategy mechanism (a posted price, which each party can accept or refuse, and they trade if both accept). For other priors, she can do strictly better. Yamashita also considers an auction setting with interdependent values, and shows that maxmin expected revenue may be achieved by a second-price auction with reserve.

In situations when one can do robustly better than dominant-strategy mechanisms, as arise in Börgers & Smith (2012) or Yamashita (2015a), optimal mechanisms are usually not known. Instead of trying to solve this hard optimization problem, the designer might look within a particular, interpretable class of mechanisms that is less restrictive than dominant-strategy. One such proposal is by Börgers & Li (2017), who explore mechanisms in which an agent's optimal strategy depends on his preferences and first-order beliefs about others' preferences, but not on higher-order beliefs. This includes, for example, trading mechanisms in which one agent makes a take-it-or-leave-it price offer to another.

There have been a few other lines of work that could also be classified under "robustness to strategic behavior," that adopt various novel solution concepts to describe how agents behave. For example, Li (2017) argues that practical mechanisms should satisfy a criterion even stronger than strategy-proofness, namely **obvious strategy-proofness**: they should be implemented by an extensive form in which, at every stage, an agent who follows his "obviously dominant" strategy is guaranteed a better outcome than *any* outcome he could get if he deviates. This way, each agent can see that his strategy is optimal without needing to think about how his outcome is determined by other agents' strategies. An ascending auction is obviously strategy-proof, while the second-price sealed-bid auction (which traditional theory holds to be equivalent) is strategy-proof but not obviously so. Eliaz (2002) proposes that mechanisms should ensure good outcomes when up to k agents are "faulty" and behave in a totally unpredictable manner; he studies a version of the Maskin (1999) implementation problem under this model of behavior. There has also been some work using models of learning by boundedly rational agents, aiming to design mechanisms such that repeated play converges to desirable outcomes in the long run. Healy & Mathevet (2012) consider mechanisms in which the best-reply mapping is a contraction, which ensures convergence under various dynamics. Sandholm (2002, 2005, 2007) studies Pigouvian-style congestion pricing mechanisms and shows that various natural dynamics converge to socially efficient outcomes.

obviously strategy-proof mechanism: one where each agent has a strategy that always guarantees a better outcome than any outcome reachable under deviation

2.4. Robustness to Distributions

Traditional Bayesian design problems put a prior distribution on unknown features of the environment (such as agents' preferences), and maximize the expectation of some objective, such as profit. It can be natural to ask what happens if the designer has only partial information about the distribution, and wishes to maximize a guarantee under this par-

of the implementation problem, including some discussion of various robustness issues, although it does not emphasize undominated-strategy implementation.

tial information. This is an especially natural question if the preference types are high-dimensional or otherwise complex objects and formulating a prior is difficult. Note also that this is distinct from the question discussed in Subsection 2.2, where the designer's prior over preferences was fixed, and the uncertainty was only about agents' beliefs about each other.

A natural starting point is to consider the simplest standard mechanism design problem: a monopolist selling a single object, to a buyer with unknown value v drawn from some distribution, trying to maximize expected profit. For example, what happens if the seller instead does not know the distribution, but only knows the mean and an upper bound on v, and wishes to design a mechanism to maximize expected profit in the worst case over all distributions consistent with this knowledge? Intuitively, the seller would want to randomize the price in order to hedge the uncertainty. A given buyer type then gets the good with probability strictly between 0 and 1; thus, randomizing prices is equivalent to screening buyers by offering a menu of probabilities q of receiving the good, and specifying a price $p(q)$ for each choice of q. This randomization is in contrast to the case of a known distribution, where it is always optimal to set a single, deterministic price (Riley & Zeckhauser 1983). Carrasco, Luz, Kos, Messner, Monteiro & Moreira (2018a) explicitly derive the optimal distribution over prices. (They also consider a generalization where multiple moments of the value distribution are known, although in this case the optimal mechanism cannot be given explicitly.)

Many variants of this problem quickly present themselves. Carrasco, Luz, Monteiro & Moreira (2018b) consider a version where the monopolist can sell continuous quantities and the agent has nonlinear preferences, and characterize the maxmin-optimal mechanism by an ODE. Bergemann & Schlag (2008) consider a totally prior-free model: the seller only knows that the buyer's value lies in $[0, 1]$. Here the maxmin expected profit objective is uninteresting (the worst case is simply that the buyer's value is 0 for sure), but they instead consider minmax regret — that is, pricing so as to minimize the worst-case value of the difference between realized profit and the profit the seller could have gotten if she knew the buyer's true value. Again, the optimum involves randomizing prices and they derive the relevant distribution. (See also Bergemann & Schlag 2011.) Auster (2018) considers a monopolist problem with interdependent values: the seller's cost of providing the good also depends on the buyer's type. In this model, maxmin expected profit over all distributions (equivalently, over all possible buyer types) is a nontrivial problem, and she characterizes the solution to this, as well as versions with less extreme uncertainty.

Although models of this sort are natural to write down, it is not always clear what lessons one can expect to learn from them that would not equally well arise in the corresponding fully-Bayesian model (aside from the idea that uncertainty can be hedged by randomization, which is just a fundamental property of maxmin models, see e.g. Gilboa & Schmeidler (1989)). Ideally, one would like to be able to give some economic interpretation to the specific form of the maxmin-optimal mechanism. One study that pursues such a concrete interpretation is the principal-agent model of Carroll & Meng (2016b). They examine the idea that linear contracts give the same incremental incentive for effort at every point. In their model, the parties contract on output, which equals (one-dimensional) effort plus a random noise term. The agent first observes the noise, then chooses effort; this is thus a "false moral hazard" model. (It also is isomorphic to a version of the Laffont & Tirole (1986) procurement model.) The principal knows the agent's effort cost function, but does not know the distribution of noise, only its mean. Because a linear contract always induces the same effort regardless of the noise realization, the principal's expected profit depends on the noise distribution only through its mean, which makes such contracts a natural candidate for the maxmin optimum. In fact, the optimal contract is indeed linear, except for a flat part at the bottom where a limited liability constraint binds.

A different class of models that naturally lends itself to uncertainty about distributions is models with multidimensional types — both because the assumption of a fully-specified prior distribution can be especially strained if the type space is high-dimensional, and because standard Bayesian models tend to lead to overly complicated predictions. For example, consider the natural multidimensional generalization of the monopolist problem: the monopolist sells K goods, to a buyer whose values for the goods are unknown (for simplicity, the buyer's preferences are additive across the goods). Even if the values for the goods are independently distributed, the optimal selling mechanism involves bundling the goods, and can even involve offering a menu of infinitely many probabilistic bundles at different prices; see Daskalakis, Deckelbaum & Tzamos (2013, 2017). Carroll (2017) considers the following robust variant: instead of assuming a joint prior distribution over the values for the goods, assume the seller only knows the marginal distribution on each good separately. The seller wishes to maximize expected profit, in the worst case over joint distributions that are consistent with the known marginals. A natural candidate for the optimal mechanism is to sell each good separately, since the total profit then does not depend on the details of the joint distribution. Carroll shows this is indeed the maxmin optimum. Moreover, the result generalizes considerably, to any situation where an agent is to be assigned a K-dimensional allocation based on K corresponding dimensions of private information, and preferences are quasi-linear and separable across the dimensions. The proof uses a somewhat involved construction of a worst-case joint distribution. Gravin & Lu (2018) extend the monopolist result to allow for a budget-constrained buyer.

Dworczak (2017) uses a distributional-robustness argument to motivate a particular class of mechanisms. He studies a setting in which a designer provides a mechanism, agents participate in the mechanism, and then later they may participate in some further interaction that is outside the designer's control, but whose outcome depends on information revealed by the mechanism. Identifying optimal mechanisms is challenging, but he restricts to a tractable class of mechanisms based on "cutoff" allocation rules. Such rules are characterized by the following property: they can always be implemented by some payment rule, regardless of the distribution of types and the nature of the post-mechanism interaction (although the specific payment rule does depend on these data). He offers some suggestive arguments for why this property may be desirable.

Most of the works above considered a single agent, with type drawn from an unknown distribution. When there are multiple agents, it is natural to try to learn the distribution. For example, one could look at samples — either previous or simultaneous participants in the mechanism. In fact, this idea would apply also in a Bayesian model, with a prior distribution over distributions. Segal (2003) studies precisely such a Bayesian model, for a seller selling identical goods with a nonlinear production cost. He derives the exactly-optimal mechanism, and compares the rate of convergence with many buyers (toward the profit that would be attainable if the distribution were known in advance) against some intuitive prior-free learning approaches, such as sequential experimentation with prices, or asking a subset of buyers to report their values and using the estimated demand curve to set a price for remaining buyers.

In such sampling environments, assuming that one is unwilling or unable to commit to a prior over distributions, a natural maxmin-style goal is to come close to the performance that would have been attainable if the distribution had been known, and do so uniformly over a wide range of distributions. This has been explored in some recent literature (e.g. Cole & Roughgarden (2014), Huang, Mansour & Roughgarden (2015)). Finding exact-maxmin optimal mechanisms seems to be intractable, so the focus is instead on achieving optimal or near-optimal asymptotic convergence rates. This involves technical constructions to efficiently hedge against the possibility of drawing samples that are unrepresentative of the true distribution. Morgenstern & Roughgarden (2015) consider a designer who is restricted to a particular class of auction formats, and relate the attainable convergence rates to a measure of complexity of the class.

Another approach to learn the distribution — and one that applies even with a small number of agents — is to simply ask the agents about it. If the agents' beliefs are assumed to come from a common prior, then, once again, the general-purpose textbook answer is to ask the agents to report the prior, and punish them all if the reports disagree. But one would like mechanisms that make less reliance on agents' precise knowledge. Caillaud & Robert (2005) consider the Myerson (1981) single-good auction setting, maintaining the assumption of independent private values, and show how the optimal auction can be implemented without knowing the distribution by effectively having bidders propose reserve prices for other bidders. Relatedly, Brooks (2013) considers a single-good auction in which buyers' types need not be independently distributed, and the only thing known about the joint distribution is a bound on the ratio of the highest possible value to the to the expected surplus available. The seller maximizes the worst-case ratio of expected revenue to expected feasible surplus. The optimum is achieved by a "surveying and selling" mechanism, which asks each buyer to report two things: his value for the good, and his belief about the distribution of the highest of others' values. The bidder reporting the highest value is then offered the good at a price based on the distribution reported by another bidder. Additional incentives can be provided to induce truthful reporting of beliefs.

collusion: multiple agents coordinating behavior in a mechanism so as to achieve a jointly better outcome

renegotiation (reallocation): agents agreeing to change the outcome specified by the mechanism, to another outcome that is mutually preferred

2.5. Robustness to Interactions among Agents

In the usual approach to mechanism design, when a principal offers a mechanism to a group of agents, it is implicitly assumed that the agents interact only through the mechanism. But there are at least two ways this assumption could be violated. Agents may be able to **collude** in the mechanism, perhaps after first exchanging some information in order to decide how to play. Collusion is a major concern in auction practice. And they may also be able to **renegotiate** (or **reallocate**) ex post. For example, if the mechanism sells goods to some of the agents, they may then resell the goods among themselves.

It is hard to find a generally satisfactory way of modeling how collusion might take place. The social choice literature has dealt with the collusion issue by adopting an agnostic and strong requirement, in the same spirit as strategy-proofness, namely **group strategy-proofness**. This criterion is quite demanding, but in numerous environments there are interesting mechanisms that satisfy it (e.g. Dubins & Freedman 1981, Bird 1984), and it can even be implied by individual strategy-proofness in some situations (see Le Breton & Zaporozhets 2009 and Barberà et al. 2010). Bierbrauer & Hellwig (2016) consider a variant criterion that reduces the scope for coalitional deviations by requiring them to not be vulnerable to further deviation by subcoalitions, but also enhances it by allowing for asymmetric information.

group strategy-proof: no coalition of agents can ever jointly misreport preferences so as to make each of them better off

In settings with monetary transfers, such as auctions, one might imagine that members of a collusive coalition would make side payments among themselves. Then group strategy-proofness is not enough, since a coalition could deviate to increase its total payoff, harming some members but then compensating them with side payments. In these settings, we might instead expect coalitions to coordinate to maximize the sum of members' payoffs. Chen & Micali (2012) consider a model where players are grouped into coalitions and the grouping is unknown to the designer. They formulate a version of the dominant-strategy property where each agent is asked to report both his individual preferences and the set of other agents he is colluding with, and propose a single-good auction that is collusively strategy-proof, i.e. no coalition can benefit by jointly misreporting. Chen & Micali (2009) study a combinatorial auction environment and adopt a much more agnostic model of coalition behavior, and achieve nontrivial revenue guarantees expressed in terms of the best-informed non-colluding player's knowledge of others' values.

If we return to imperfect-information settings, and are willing to assume a common prior (shared by the designer and the agents), then Che & Kim (2006) give a strong positive result on collusion-proofness. They consider an environment with quasilinear preferences, and they allow a quite general class of collusive protocols, but assume (following Laffont & Martimort 1997) that collusion is limited by the same informational constraints as the original designer faces. That is: they consider procedures in which, once the mechanism has been proposed and agreed to, the agents can formulate a collusive side contract, which may potentially involve both manipulating the mechanism and reallocating afterward; but they can lie about their types to each other during side contracting, just as they can in the original mechanism. They show that any mechanism that could be implemented without collusion can be made resistant to all collusion procedures of this sort, by an appropriate adjustment of the payment functions. In effect, their construction "sells" control of the mechanism to the agents collectively. A companion paper, Che & Kim (2009), considers an alternative timing in which side contracting happens before the agents agree to participate in the mechanism. This opens up new possibilities for collusion since the side contract can sometimes instruct agents not to participate. They focus on a single-good auction, and show that under many circumstances, it is again possible to implement the Myerson (1981) optimal auction in a way that makes it collusion-proof.

The positive results of Che & Kim (2006, 2009) may seem unrealistically optimistic. One view is that they show what needs to be added to the model so that the possibility of collusion has bite. For example, these models make the usual strong assumptions of a common prior among the designer and all agents. If one added a concern about collusion to a belief-robust model as in Chung & Ely (2007), the collusion constraint would be binding.[9]

A different weakness of the mechanisms in Che & Kim (2006, 2009) is that they are not ex post individually rational — agents sometimes end up with a negative payoff. Motivated by this, Che, Condorelli & Kim (2016) consider a "winner-payable" class of auctions, roughly defined by the property that any bidder can potentially win the object with all other bidders paying nothing. They also adopt a weaker model of collusion in which bidders cannot make side payments to each other, so that the distribution of payments across individual bidders matters. They characterize the optimal collusion-proof auction within this class. In general, it is strictly worse than the seller could do without collusion.

[9] A worst case is when the agents all know each other's preferences. If they also can freely reallocate and make side transfers, then they effectively combine into a single agent.

Moving on from collusion to pure renegotiation, there are again challenges as to what the right model should be. One traditional approach is to treat the renegotiation procedure as a "black box" — an exogenously given function $h(x,\theta)$ that describes what outcome would arise if the agents' type profile were θ and allocation x were specified by the mechanism. Even when a designer looks for a mechanism that always specifies a Pareto-efficient outcome, the possibility of renegotiation can impose additional constraints, because an agent could strategically deviate in the mechanism, obtain an inefficient outcome, and then renegotiate it. Maskin & Moore (1999) and Segal & Whinston (2002) study the problem of implementation under complete information when agents can renegotiate, modeled by the black-box approach.[10]

In reality, a designer concerned about renegotiation might not know the details of the renegotiation procedure $h(\cdot,\cdot)$. Neeman & Pavlov (2013) address this by formulating a criterion of **ex-post renegotation-proofness**, an analogue of strategy-proofness for such environments. This criterion requires that the outcome specified by the mechanism should not be vulnerable to manipulation followed by renegotiation, for any (individually rational) renegotiation procedure. They characterize mechanisms satisfying this property, in a complete-information environment, and make some inroads on an extension to incomplete information. While ex-post renegotiation-proofness provides a strong guarantee against renegotiation, it is potentially open to the "foundations" critique: the designer's concern is really with the outcome, and restrictions on the mechanism are only one means to this end. One might be able to achieve better robust guarantees by allowing mechanisms in which agents may sometimes renegotiate along the equilibrium path (and their actions in the mechanism may depend on the renegotiation procedure).

Carroll & Segal (2018) address this possibility head-on. They study a single-good auction problem, where a seller wants to maximize expected revenue; here, renegotiation consists of bidders reselling the good among themselves after the auction. The auction designer is maximizing expected revenue, so cares about resale only because the prospect of resale affects bidders' behavior in the auction itself. They assume an asymmetric prior: some bidders are "stronger," i.e. more likely to have high values for the good, than others. In such a setting, Myerson's (1981) classic optimal auction would discriminate against the stronger bidders, sometimes selling the good to a weaker bidder when a stronger bidder has a higher value. Consequently, the possibility of resale has bite.[11] If the modeler incorporates resale by specifying some particular resale game after the auction, typically the optimal auction still involves some discrimination, after which resale may occur in equilibrium. Carroll and Segal instead consider the problem of maximizing worst-case expected revenue, where the worst case is over possible resale procedures. They show that the optimum is attained by a particular "Ausubel-Cramton-Vickrey" auction (Ausubel & Cramton 2004), which sometimes withholds the good via reserve prices, but if it sells, it always sells to the highest-value

[10] Other approaches have been considered, though less relevant here. Rubinstein & Wolinsky (1992) assume that the renegotiation procedure is not known, but is certain to destroy surplus due to delay. Jackson & Palfrey (1998, 2001) consider the related topic of "voluntary implementation," which assumes that agents can veto the outcome and play the mechanism again, thereby endogenizing the outcome that actually results when the mechanism specifies a bad outcome off-path.

[11] Che and Kim's (2006) collusion-proofness construction does not apply to guard against resale, because their collusion happens before the mechanism, whereas resale occurs after the auction. The difference is that after the auction, some information about the agents' types has already been revealed (via the auction outcome).

bidder (who then does not resell). Thus, unlike the approach of Neeman & Pavlov (2013), they do not impose renegotiation-proofness a priori, but it emerges in the solution to their maxmin problem.

2.6. Local Versions of Robustness

All of the above studies considered "global" notions of robustness — in which a designer wants to ensure that a mechanism performs uniformly well in some large class of environments. One can instead study "local" versions of robustness, where a designer tailors a mechanism to some benchmark model of the environment, but wants to ensure that the mechanism still performs well if the environment is slightly misspecified (or, equivalently, if the environment later changes a little bit). In many cases, qualitative properties of the optimal mechanism are unchanged; the mechanism looks like in the benchmark model but with some adjustments. Still, it may be useful to know what form those adjustments should take. It can also be conceptually useful to go through with the exercise in order to identify situations where such small adjustments are sufficient, versus situations where the model is fundamentally unsound to local perturbations.

Madarász & Prat (2017) perform the local robustness exercise in a standard screening model. One can think, for example, of a monopolist selling several goods to buyers with some distribution of preferences. If we take the distribution as given and solve for the optimal mechanism, then profit as a function of the buyer's type is typically discontinuous, which means that profit can fall by a lot if the model is slightly misspecified. (For example, imagine that the seller has one good, and a buyer's value is predicted to be 1, 2, or 3, with probability 1/3 each. The optimal selling price is 2, yielding expected profit 4/3. But if the model is wrong and the buyer type that was predicted to have value 2 actually has value $2 - \epsilon$, this type does not buy and profit falls to 2/3.) They show that a simple fix — rebating a small fraction of profit to the buyer — makes the mechanism locally robust, ensuring at most a small drop in profits relative to the benchmark for any true distribution that is close to the benchmark model. If the amount of misspecification is ϵ, the fraction of profit that should be rebated is on the order of $\sqrt{\epsilon}$, and the worst-case loss in profit is also on the order of $\sqrt{\epsilon}$. Carroll & Meng (2016a) give an analogous result for a moral hazard model, as well as a much more general class of mechanism design problems. In the moral hazard model, they also show that the construction is asymptotically optimal (to within a constant factor), i.e. there is no way to guarantee less than $O(\sqrt{\epsilon})$ shortfall relative to the benchmark for small ϵ.

These works formalize local robustness by taking a maxmin over an ϵ-sized neighborhood of the benchmark model. Another approach that is popular elsewhere in economics is the multiplier preferences of Hansen & Sargent (2001), in which the amount of performance degradation that is tolerated increases in a continuous way as one considers alternative environments farther from the benchmark. Miao & Rivera (2016) take this approach to study a locally robust version of a dynamic moral hazard contracting problem, based on the financial contracting models of DeMarzo & Sannikov (2006) and Biais, Mariotti, Plantin & Rochet (2007). They interpret their model as a study of how financial contracts are affected by ambiguity aversion, and relate it to empirical evidence on variation in asset prices across firms, especially patterns in the equity premium (by contrast, the non-robust benchmark version of the model does not generate any equity premium). Thus the robust contracting problem serves as a modeling tool to study other phenomena, in line with item 4 from our taxonomy back in Subsection 1.2.

One area of mechanism design where local robustness can make a drastic difference is in implementation under complete information. Recall the example from Subsection 1.1.3, of a buyer and seller, trading a good of low or high quality. We saw that the Moore & Repullo (1988) mechanism implemented the desired outcome in subgame-perfect equilibrium, but it was sensitive to the complete-information assumption, and could break down under arbitrarily small amounts of incomplete information. The key insight is that when a small amount of incomplete information is introduced, reaching a part of the game tree that was previously off-equilibrium-path is now a very informative event, and this can discontinuously change predicted behavior. In fact, Aghion, Fudenberg, Holden, Kunimoto & Tercieux (2012) showed that if we require outcomes to be robust to an arbitrarily small amount of incomplete information, *no* dynamic mechanism can implement an outcome that violates Maskin monotonicity (as in our example). Chung & Ely (2003) also gave an analogous result in a static setting, with a different solution concept.

Oury & Tercieux (2012) return to static settings and consider the problem of requiring only partial implementation (i.e. only some equilibrium, not necessarily all equilibria, should give the desired outcome). Thus, in the buyer-seller example, we could go back to the simple mechanism of asking both parties to report and forbidding trade if they disagree. Oury and Tercieux observe that if we allow belief perturbations that are not consistent with a common prior (something that the studies in the preceding paragraph did not allow), then even the simple mechanism can fail. Specifically, they consider perturbations in which preferences are (almost) mutually certain to Nth order for large N but not common knowledge, as in the "email game" (Rubinstein 1989). They show that requiring robustness to such perturbations again limits the implementable social choice functions to those satisfying Maskin monotonicity.

2.7. Robustness of Standard Mechanisms

There is a body of literature that could be classified under the heading of "robustness" and warrants brief mention here, that does not consider design questions, but rather studies ways in which "standard" mechanisms perform well across a wide range of environments, often approaching first-best efficiency when there are many agents. This includes the Walrasian mechanism for exchange economies, and variants such as double auctions. See Rustichini, Satterthwaite & Williams (1994), Jackson & Manelli (1997), Cripps & Swinkels (2006), Reny & Perry (2006). At a more general level, Jackson & Kremer (2007) show that guaranteeing approximate incentives for truth-telling under unknown type distributions is closely linked to the property of envy-freeness. Azevedo & Budish (2018) argue that a similar criterion of approximate strategy-proofness in large markets helps to explain a range of observations about mechanisms that do or do not seem to perform well in practice.

3. DISCUSSION

This is a natural spot to give some reflections — to see what we have learned from the small but rather diffuse body of work so far on robustness in mechanism design, and to speculate on what might be the most productive aims for future work, both for individual contributions and for the progress of the field as a whole. These comments will naturally be more subjective than the summaries above.

3.1. Robustness, Detail, and Simplicity

In the first two motivating examples from the introduction, the optimal mechanism from the traditional theory was unreasonably tailored to detailed assumptions about the environment, and this was a clue that incorporating robustness considerations into the model might improve realism. More generally, there is an ethos in much of the mechanism design community that realistic mechanisms should not be finely tuned to parametric assumptions, such as probability distributions of values or functional forms of preferences: that they should be, in a word, "detail-free." This view is sometimes referred to as the "Wilson doctrine" or "Wilson critique" — although the attribution is murky.[12] Related to detail-freeness is the even more nebulous concept of "simplicity," also sometimes invoked as a desideratum.

But robustness is a distinct concept from detail-freeness or simplicity. Maxmin problems can sometimes have very complicated solutions; and as anyone who's tried to read a software license or cell-phone contract knows, in practice contracts are often made robust by tediously listing imaginable contingencies one by one. Conversely, classical Bayesian models can deliver appealingly simple solutions but that rely on strong assumptions (such as the d'Aspremont & Gérard-Varet (1979) expected externality mechanism, which leans on the common-prior assumption). Börgers (2015, Chapter 10) and Chung & Ely (2007) discuss the distinctions in more detail. In summary, there is no necessary *logical* relation between robustness and these other concepts. There can, however, be a *methodological* relation: observing how one model delivers complicated or sensitive predictions can suggest a way to write a robust model that makes better predictions. Much of the research surveyed here has described robust models that make simple or largely detail-free predictions, but rather than being evidence of some sublime connection between these properties, this might just be because economic theory generally is more likely to be insightful — and, therefore, publishable — when it delivers simple and understandable results.

3.2. The Role of Robust Models

Evidently, robust models of mechanism design can serve a variety of conceptual purposes, just as with traditional, fully-Bayesian models. There also are a number of different reasons one might specifically study a robust model: we may find the prediction from a traditional Bayesian model unrealistically sensitive to details; we may simply find it mathematically or computationally intractable, and hope that writing down a parallel model with a different objective will enable more progress; we may not even know how to specify the Bayesian model (in practice, how can one formulate a probabilistic prior over such abstract things as higher-order beliefs, or resale procedures?). Or we may have no specific objection to a Bayesian model, but simply wish to consider multiple models to get a broader range of perspectives. Of course, in each of these situations, the robust model may or may not actually do any better. Such models simply provide one more set of tools to try out.

Not long ago, mainstream economic modeling followed a fairly strict orthodoxy of expected-utility maximization. These days the culture is becoming more pluralistic, due

[12] This name appears e.g. in Satterthwaite & Williams (2002), Maskin (2003), Baliga & Vohra (2003), Dhangwatnotai, Roughgarden & Yan (2015). It is often backed up with a quote from Wilson (1987) saying that game-theoretic analysis of a given mechanism should not rely excessively on common-knowledge assumptions. But this is logically distinct from detail-freeness as above. In personal communication, Robert Wilson has expressed that he agrees in spirit with the "Wilson doctrine," but is surprised to receive credit for it.

partly to the success of behavioral economics and partly to a more general shift away from literalism in interpreting economic models. One does still sometimes hear the objection to (say) maxmin models, that real-life decision-makers rarely if ever maximize extreme worst-case objectives. In this author's view, this is not a serious objection, since it is also rare that decision-makers have fully-specified priors and maximize expected utility. Reality is somewhere in the hard-to-model space in between, and both of the extremes provide feasible modeling approaches that can deliver some useful insights. This is all the more true in mechanism design, where we can interpret the modeler's choice of objective not necessarily as the maximization problem faced by an actual decision-maker, but rather as a principled way of studying mathematical properties of certain mechanisms. (Recall the discussion from Subsection 2.1.)

That said, non-Bayesian models do face some extra hurdles, particularly if they are to be incorporated into larger game-theoretic settings. For one, if we wish to write equilibrium models in which multiple interacting players face maxmin-style uncertainty, this demands a more literal interpretation of the maxmin objective. Nash equilibrium implicitly presumes that each player is certain of what the other player is doing, so, in effect, each player must be uncertain about the environment yet certain that the opponent shares his uncertainty — a combination of assumptions that may strain our credulity. Another challenge is that trying to write dynamic models with non-Bayesian decision makers leads to well-known problems of dynamic inconsistency except in special cases (e.g. Epstein & Schneider 2003). This may be one reason why there has been relatively little work so far on robust mechanism design in dynamic settings.

3.3. Effective Robust Modeling

For the theorist interested in contributing to this area, what is a productive way to get started? The following general recipe seems to be one possible route; many of the sudies surveyed above can be cast as instances of it.

1. Begin with some classical mechanism design setting — one for which the "standard" prescription for the optimal mechanism seems unrealistic (or where the Bayesian problem is hard to even write down).
2. Write down a mechanism (or a small parameterized family of mechanisms) that seems like a reasonable one to use.
3. Write down some intuitive argument for why the mechanism performs well across a range of possible environments.
4. Translate the previous step into a robust optimization problem, such as a maxmin problem, for which the proposed mechanism is a natural candidate solution.
5. Solve the problem. If the proposed mechanism is in fact the solution, this provides a formalization of its robustness. If not — some other mechanism performs robustly better — then so much the merrier; we have learned something new.

Of course, this recipe will not always succeed. Nonetheless, it seems to be a productive route to generate insights. In interesting cases, typically, the maxmin criterion is essential to the analysis, in the sense that one would not simply get the same results by guessing the worst-case environment in advance and solving the corresponding Bayesian problem: for example, because the worst case is hard to guess, or because the corresponding Bayesian problem has many optimal mechanisms and the worst-case criterion helps select among them. Or, the worst-case environment may simply be one that is hard to motivate on a priori grounds, and better motivated via robustness considerations.

As in other kinds of economic models, there is flexibility in writing down the model. This flexibility arises not only in writing down the class of environments to consider. It also arises in the choice of objective to optimize. Most of the work considered above looked at maxmin objectives, but as we have seen, some instead looked at minimizing regret relative to some benchmark (and the choice of benchmark is flexible), or even some other criteria. In this author's view, the right choice of criterion is whichever one expresses useful insights; this can vary from one application to the next.[13]

3.4. The Future of Robustness in Mechanism Design

What will be the lasting lessons from this body of work on robustness? In what directions should the field develop to maximize its impact on economics, or even on incentive design in practice? Here are a few brief speculations.

Keeping in mind the taxonomy of lessons from mechanism design from Subsection 1.2, there are a few different ways that lasting contributions might emerge. One possibility is that the field may develop some basic modeling tools that become widely used — just as a few models from classical mechanism design, such as the Holmström (1979) and Grossman & Hart (1983) model of moral hazard or the Myerson (1981) model of optimal auctions, have become central in present-day economic theory. It seems likely that the strongest contenders for such central tools will be ones that are portable and tractable, and especially so if they can provide tractability for studying problems that are difficult to solve using traditional Bayesian models. (Indeed, such tools carry a mixed blessing, since they may end up often being blindly used in applications that do not fit the modeling assumptions.) At present, the state of the literature is rather diffuse and it seems hard to identify such a core set of models. But as the field grows organically and matures, it will likely become more apparent which tools can be used repeatedly.

Another possible winning scenario for the field would be finding at least one "killer app" — a practical incentive problem for which the robust approach leads to new and useful mechanisms. A weakness of the field at present is that, very often, the analyses either (a) are fairly technically involved, yet end up just giving stronger foundations to mechanisms that were already being advocated or used anyway; or (b) suggest some improvements over existing mechanisms, yet without identifying new *optimal* mechanisms or, more generally, crisp lessons about what mechanisms should be used and why.

Outcome (a) is, again, not necessarily a problem at an early stage. One view is that the

[13] Börgers (2017) proposes to refine the maxmin criterion, by observing that some mechanisms that are maxmin-optimal may nonetheless be weakly dominated: one can attach bells and whistles to make them perform better in some non-worst-case environments. This observation indicates that the simpler mechanism isn't really selected as optimal. That said, in many cases, one arguably obtains useful insights from studying the simpler mechanism and showing that it satisfies the (weaker) maxmin criterion.

A related idea that one sometimes hears is that the designer should just get the agents to fully report the unknown environment (common prior distribution, resale procedure, etc.) and then run the optimal mechanism for that environment. As a "solution" to the designer's problem, this is of course rather unenlightening; nor does it seem realistic as a literal prescription. But a theme behind some work (e.g. Börgers & Smith 2012, 2014, Brooks 2013) is that some movement in this direction may indeed be possible without abandoning practicality.

current state of the field is one of developing modeling tools, and tools should be beta-tested by seeing whether their results agree with existing solutions and intuitions, before later taking them to new problems for which no such solutions exist. Outcome (b), too, is not necessarily a bad one insofar as it demonstrates that robust models can make a difference. And being able to exactly optimize is not really necessary at least when the goal is practical design. But for the tools to be widely useful, we eventually need to go beyond scattered examples and develop systematic theory, if not to find optimal mechanisms in practice, then at least to have principled ways of approaching problems for which optimization is elusive.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This survey has benefited from conversations and comments from (in random order) Brad Larsen, Richard Holden, Matt Jackson, Stephen Morris, Tilman Börgers, Shengwu Li, Alex Wolitzky, Dirk Bergemann, and Ben Brooks. Weixin Chen provided excellent research assistance. The author is supported by a Sloan Research Fellowship. Much of this writing was done during visits to the Cowles Foundation at Yale and to the Research School of Economics at the Australian National University, and the author gratefully acknowledges their hospitality.

LITERATURE CITED

Abreu D, Matsushima H. 1992a. Virtual implementation in iteratively undominated strategies: Complete information. *Econometrica* 60(5):993–1008

Abreu D, Matsushima H. 1992b. A response to Glazer and Rosenthal. *Econometrica* 60(6):1439–1442

Abreu D, Matsushima H. 1994. Exact implementation. *J. Econ. Theory* 64(1):1–19

Aghion P, Fudenberg D, Holden R, Kunimoto T, Tercieux O. 2012. Subgame-perfect implementation under information perturbations. *Q. J. Econ.* 127(4):1843–1881

Auster S. 2018. Robust contracting under common value uncertainty. *Theor. Econ.* 13(1):175–204

Ausubel LM, Cramton P. 2004. Vickrey auctions with reserve pricing. *Econ. Theory* 23(3):493–505

Azevedo EM, Budish E. 2018. *Strategy-proofness in the large. Rev. Econ. Stud. Forthcoming*

Babaioff M, Lavi R, Pavlov E. 2009. Single-value combinatorial auctions and algorithmic implementation in undominated strategies. *J. ACM* 56(1):#5

Baliga S, Vohra R. 2003. Market research and market design. *Advances Theor. Econ.* 3(1):#5

Barberà S. 2011. Strategy-proof social choice. In *Handbook of Social Choice and Welfare, vol. 2*, eds. KJ Arrow, A Sen, K Suzumura. Amsterdam: North-Holland, 731–831

Barberà S, Berga D, Moreno B. 2010. Individual versus group strategy-proofness: When do they coincide? *J. Econ. Theory* 145(4):1648–1674

Battigalli P, Siniscalchi M. 2003. Rationalizable bidding in first-price auctions. *Games Econ. Behav.* 45(1):38–72

Bergemann D, Brooks B, Morris S. 2016. *Informationally robust optimal auction design.* Unpublished manuscript, Univ. Chicago, Chicago, IL

Bergemann D, Brooks B, Morris S. 2017. First-price auctions with general information structures: Implications for bidding and revenue. *Econometrica* 85(1):107–143

Bergemann D, Brooks B, Morris S. 2018. *Revenue guarantee equivalence.* Unpublished manuscript, University of Chicago, Chicago, IL

Bergemann D, Morris S. 2005. Robust mechanism design. *Econometrica* 73(6):1771–1813

Bergemann D, Morris S. 2009a. Robust implementation in direct mechanisms. *Rev. Econ. Stud.* 76(4):1175–1204

Bergemann D, Morris S. 2009b. Robust virtual implementation. *Theor. Econ.* 4(1):45–88

Bergemann D, Morris S. 2011. Robust implementation in general mechanisms. *Games Econ. Behav.* 71(2):261–281

Bergemann D, Morris S. 2017. *Information design: A unified perspective.* Unpublished manuscript, Yale Univ., New Haven, CT

Bergemann D, Morris S, Tercieux O. 2011. Rationalizable implementation. *J. Econ. Theory* 146(3):1253–1274

Bergemann D, Schlag KH. 2008. Pricing without priors. *J. Eur. Econ. Assoc.* 6(2/3):560–569

Bergemann D, Schlag KH. 2011. Robust monopoly pricing. *J. Econ. Theory* 146(6):2527–2543

Biais B, Mariotti T, Plantin G, Rochet JC. 2007. Dynamic security design: Convergence to continuous time and asset pricing implications. *Rev. Econ. Stud.* 74(2):345–390

Bierbrauer FJ, Hellwig MF. 2016. Robustly coalition-proof incentive mechanisms for public good provision are voting mechanisms and vice versa. *Rev. Econ. Stud.* 83(4):1440–1464

Bird CG. 1984. Group incentive compatibility in a market with indivisible goods. *Econ. Lett.* 14(4):309–313

Börgers T. 1991. Undominated strategies and coordination in normalform games. *Soc. Choice Welf.* 8(1):65–78

Börgers T. 2015. *An Introduction to the Theory of Mechanism Design.* New York: Oxford Univ. Press

Börgers T. 2017. (No) foundations of dominant-strategy mechanisms: A comment on Chung and Ely (2007). *Rev. Econ. Design* 21(2):73–82

Börgers T, Li J. 2017. Strategically simple mechanisms. Unpublished manuscript, Univ. Michigan, Ann Arbor, MI

Börgers T, Smith D. 2012. Robustly ranking mechanisms. *Am. Econ. Rev.* 102(3):325–329

Börgers T, Smith D. 2014. Robust mechanism design and dominant strategy voting rules. *Theor. Econ.* 9(2):339–360

Brooks B. 2013. *Surveying and selling: Belief and surplus extraction in auctions.* Unpublished manuscript, Univ. of Chicago, Chicago, IL

Brooks B, Du S. 2018. *Optimal auction design with common values: An informationally-robust approach.* Unpublished manuscript, Univ. of Chicago, Chicago, IL

Caillaud B, Robert J. 2005. Implementation of the revenue-maximizing auction by an ignorant seller. *Rev. Econ. Design* 9(2):127–143

Carrasco V, Luz VF, Kos N, Messner M, Monteiro P, Moreira H. 2018a. Optimal selling mechanisms under moment conditions. *J. Econ. Theory* Forthcoming

Carrasco V, Luz VF, Monteiro P, Moreira H. 2018b. Robust mechanisms: The curvature case. *Econ. Theory* Forthcoming

Carroll G. 2015. Robustness and linear contracts. *Am. Econ. Rev.* 105(2):536–563

Carroll G. 2016. Informationally robust trade and limits to contagion. *J. Econ. Theory* 166:334–361

Carroll G. 2017. Robustness and separation in multidimensional screening. *Econometrica* 85(2):453–488

Carroll G. 2018a. *Information games and robust trading mechanisms.* Unpublished manuscript, Stanford Univ., Stanford, CA

Carroll G. 2018b. *Robust incentives for information acquisition.* Unpublished manuscript, Stanford Univ., Stanford, CA

Carroll G, Meng D. 2016a. Locally robust contracts for moral hazard. *J. Math. Econ.* 62:36–51

Carroll G, Meng D. 2016b. Robust contracting with additive noise. *J. Econ. Theory* 166:586–604

Carroll G, Segal I. 2018. Robustly optimal auctions with unknown resale opportunities. *Rev. Econ. Stud.* Forthcoming

Chassang S. 2013. Calibrated incentive contracts. *Econometrica* 81(5):1935–1971

Che YK, Condorelli D, Kim J. 2016. *Weak cartels and collusion-proof auctions.* Unpublished manuscript, Columbia Univ., New York, NY

Che YK, Kim J. 2006. Robustly collusion-proof implementation. *Econometrica* 74(4):1063–1107

Che YK, Kim J. 2009. Optimal collusion-proof auctions. *J. Econ. Theory* 144(2):565–603

Chen J, Micali S. 2009. A new approach to auctions and resilient mechanism design, In *STOC '09 Proceedings of the forty-first annual ACM symposium on Theory of Computing,* pp. 503–512. New York: ACM

Chen J, Micali S. 2012. Collusive dominant-strategy truthfulness. *J. Econ. Theory* 147(3):1300–1312

Chen YC, Li J. 2017. *Revisiting the foundations of dominant-strategy mechanisms.* Unpublished manuscript, Nat. Univ. of Singapore, Singapore

Chung KS, Ely JC. 2003. Implementation with near-complete information. *Econometrica* 71(3):857–871

Chung KS, Ely JC. 2007. Foundations of dominant-strategy mechanisms. *Rev. Econ. Stud.* 74(2):447–476

Cole R, Roughgarden T. 2014. The sample complexity of revenue maximization, In *STOC '14 Proceedings of the forty-sixth annual ACM symposium on Theory of Computing,* pp. 243–252. New York: ACM

Crémer J, McLean RP. 1988. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica* 56(6):1247–1258

Cripps MW, Swinkels JM. 2006. Efficiency of large double auctions. *Econometrica* 74(1):47–92

Dai T, Toikka J. 2017. *Robust incentives for teams.* Unpublished manuscript, Mass. Inst. of Technology, Cambridge, MA

Daskalakis C, Deckelbaum A, Tzamos C. 2013. Mechanism design via optimal transport, In *EC '13 Proceedings of the fourteenth ACM conference on Electronic Commerce,* pp. 269–286. New York: ACM

Daskalakis C, Deckelbaum A, Tzamos C. 2017. Strong duality for a multiple-good monopolist. *Econometrica* 85(3):735–767

d'Aspremont C, Gérard-Varet LA. 1979. Incentives and incomplete information. *J. Public Econ.* 11(1):25–45

DeMarzo P, Sannikov Y. 2006. Optimal security design and dynamic capital structure in a continuous-time agency model. *J. Finance* 61(6):2681–2724

Dhangwatnotai P, Roughgarden T, Yan Q. 2015. Revenue maximization with a single sample. *Games Econ. Behav* 91:318–333

Diamond P. 1998. Managerial incentives: On the near linearity of optimal compensation. *J. Polit. Econ.* 106(6):931–957

Du S. 2018. Robust mechanisms under common valuation. *Econometrica* 86(5):1569–1588

Dubins LE, Freedman DA. 1981. Machiavelli and the Gale-Shapley algorithm. *Amer. Math. Monthly* 88(7):485–494

Dworczak P. 2017. *Mechanism design with aftermarkets: Cutoff mechanisms.* Unpublished manuscript, Northwestern Univ., Evanston, IL

Eliaz K. 2002. Fault tolerant implementation. *Rev. Econ. Stud.* 69(3):589–610

Epstein LG, Schneider M. 2003. Recursive multiple-priors. *J. Econ. Theory* 113:1–31

Frankel A. 2014. Aligned delegation. *Am. Econ. Rev.* 104(1):66–83

Garrett D. 2014. Robustness of simple menus of contracts in cost-based procurement. *Games Econ. Behav* 87:631–641

Gibbard A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41(4):587–601

Gilboa I, Schmeidler D. 1989. Maxmin expected utility with non-unique prior. *J. Math. Econ.* 18(2):141–153

Glazer J, Rosenthal RW. 1992. A note on Abreu-Matsushima mechanisms. *Econometrica* 60(6):1435–1438

Gravin N, Lu P. 2018. Separation in correlation-robust monopolist problem with budget, In *SODA '18: Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms,* pp. 2069–2080. Philadelphia, PA: SIAM

Grossman SJ, Hart OD. 1983. An analysis of the principal-agent problem. *Econometrica* 51(1):7–46

Hansen LP, Sargent TJ. 2001. Robust control and model uncertainty. *Am. Econ. Rev.* 91(2):60–66

Hartline J. 2012. Approximation in mechanism design. *Am. Econ. Rev.* 102(3):330–336

Healy PJ, Mathevet L. 2012. Designing stable mechanisms for economic environments. *Theor. Econ.* 7(3):609–661

Holmström B. 1979. Moral hazard and observability. *Bell J. Econ.* 10(1):74–91

Holmström B, Milgrom P. 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55(2):303–328

Huang Z, Mansour Y, Roughgarden T. 2015. Making the most of your samples, In *EC '15: Proceedings of the Sixteenth ACM Conference on Economics and Computation,* pp. 45–60. New York: ACM

Hurwicz L. 1972. On informationally decentralized systems. In *Decision and Organization: A Volume in Honor of Jacob Marschak*, eds. CB McGuire, R Radner. Amsterdam: North-Holland, 297–336

Hurwicz L, Shapiro L. 1978. Incentive structures maximizing residual gain under incomplete information. *Bell J. Econ.* 9(1):180–191

Hylland A. 1980. *Strategy proofness of voting procedures with lotteries as outcomes and infinite sets of strategies.* Unpublished manuscript, Harvard Univ., Cambridge, MA

Jackson MO. 1992. Implementation in undominated strategies: A look at bounded mechanisms. *Rev. Econ. Stud.* 59(4):757–775

Jackson MO. 2001. A crash course in implementation theory. *Soc. Choice Welf.* 18(4):655–708

Jackson MO, Kremer I. 2007. Envy-freeness and implementation in large economies. *Rev. Econ. Design* 11(3):185–198

Jackson MO, Manelli AM. 1997. Approximately competitive equilibria in large finite economies. *J. Econ. Theory* 77(2):354–376

Jackson MO, Palfrey TR. 1998. Efficient and voluntary implementation in markets with repeated pairwise bargaining. *Econometrica* 66(6):1353–1388

Jackson MO, Palfrey TR. 2001. Voluntary implementation. *J. Econ. Theory* 98(1):1–25

Jehiel P, Meyer-ter-Vehn M, Moldovanu B, Zame WR. 2006. The limits of ex post implementation. *Econometrica* 74(3):585–610

Laffont JJ, Martimort D. 1997. Collusion under asymmetric information. *Econometrica* 65(4):875–911

Laffont JJ, Tirole J. 1986. Using cost observation to regulate firms. *J. Polit. Econ.* 94(3):614–641

Larsen BJ. 2018. *The efficiency of real-world bargaining: Evidence from wholesale used-auto auctions.* Unpublished manuscript, Stanford Univ., Stanford, CA

Le Breton M, Zaporozhets V. 2009. On the equivalence of coalitional and individual strategy-proofness properties. *Soc. Choice Welf.* 33(2):287–309

Li S. 2017. Obviously strategy-proof mechanisms. *Am. Econ. Rev.* 107(11):3257–3287

Madarász K, Prat A. 2017. Sellers with misspecified models. *Rev. Econ. Stud.* 84(2):790–815

Marku K, Ocampo Díaz S. 2017. *Robust contracts in common agency.* Unpublished manuscript, Univ. Minnesota, Minneapolis, MN

Maskin E. 1999. Nash equilibrium and welfare optimality. *Rev. Econ. Stud.* 66(1):23–38

Maskin E. 2003. Auctions and efficiency. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, eds. M Dewatripont, LP Hansen, SJ Turnovsky. Cambridge, UK: Cambridge Univ. Press, 1–24

Maskin E, Moore J. 1999. Implementation and renegotiation. *Rev. Econ. Stud.* 66(1):39–56

McAfee RP, Reny PJ. 1992. Correlated information and mechanism design. *Econometrica* 60(2):395–421

Miao J, Rivera A. 2016. Robust contracts in continuous time. *Econometrica* 84(4):1405–1440

Mookherjee D. 2006. Decentralization, hierarchies, and incentives: A mechanism design perspective. *J. Econ. Lit.* 44(2):367–390

Moore J, Repullo R. 1988. Subgame perfect implementation. *Econometrica* 56(5):1191–1220

Morgenstern J, Roughgarden T. 2015. The pseudo-dimension of near-optimal auctions, In *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems* vol. 1, pp. 136–144. Cambridge, MA: MIT Press

Myerson RB. 1981. Optimal auction design. *Math. Oper. Res.* 6(1):58–73

Myerson RB, Satterthwaite MA. 1983. Efficient mechanisms for bilateral trading. *J. Econ. Theory* 29(2):265–281

Neeman Z. 2003. The effectiveness of english auctions. *Games Econ. Behav.* 43(2):214–238

Neeman Z, Pavlov G. 2013. Ex post renegotiation-proof mechanism design. *J. Econ. Theory* 148(2):473–501

Nisan N, Roughgarden T, Tardos E, Vazirani VV. 2007. *Algorithmic Game Theory*. Cambridge, UK: Cambridge Univ. Press

Oury M, Tercieux O. 2012. Continuous implementation. *Econometrica* 80(4):1605–1637

Reny PJ, Perry M. 2006. Toward a strategic foundation for rational expectations equilibrium. *Econometrica* 74(5):1231–1269

Riley J, Zeckhauser R. 1983. Optimal selling strategies: When to haggle, when to hold firm. *Q. J. Econ.* 98(2):267–289

Roth AE, Peranson E. 1999. The redesign of the matching market for American physicians: Some engineering aspects of economic design. *Am. Econ. Rev.* 89(4):748–780

Roughgarden T, Talgam-Cohen I. 2019. Approximately optimal mechanism design. *Ann. Rev. Econ.*

Rubinstein A. 1989. The electronic mail game: Strategic behavior under 'almost common knowledge'. *Am. Econ. Rev.* 79(3):385–391

Rubinstein A, Wolinsky A. 1992. Renegotiation-proof implementation and time preferences. *Am. Econ. Rev.* 82(3):600–614

Rustichini A, Satterthwaite MA, Williams SR. 1994. Convergence to efficiency in a simple market with incomplete information. *Econometrica* 62(2):1041–1063

Sandholm WH. 2002. Evolutionary implementation and congestion pricing. *Rev. Econ. Stud.* 69(3):667–689

Sandholm WH. 2005. Negative externalities and evolutionary implementation. *Rev. Econ. Stud.* 72(3):885–915

Sandholm WH. 2007. Pigouvian pricing and stochastic evolutionary implementation. *J. Econ. Theory* 132(1):367–382

Satterthwaite MA. 1975. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *J. Econ. Theory* 10(2):187–217

Satterthwaite MA, Williams SR. 2002. The optimality of a simple market mechanism. *Econometrica* 70(5):1841–1863

Segal I. 2003. Optimal pricing mechanisms with unknown demand. *Am. Econ. Rev.* 93(3):509–529

Segal I, Whinston MD. 2002. The Mirrlees approach to mechanism design with renegotiation (with applications to hold-up and risk sharing). *Econometrica* 70(1):1–45

Sprumont Y. 1995. Strategyproof collective choice in economic and political environments. *Can. J. Econ.* 28(1):68–107

Wilson R. 1987. Game-theoretic approaches to trading processes. In *Advances in Economic Theory: Fifth World Congress*, ed. TF Bewley. Cambridge, UK: Cambridge Univ. Press, 33–77

Yamashita T. 2015a. Implementation in weakly undominated strategies: Optimality of second-price auction and posted-price mechanism. *Rev. Econ. Stud.* 82(3):1223–1246

Yamashita T. 2015b. Strategic and structural uncertainty in robust implementation. *J. Econ. The-*

ory 159:267–279

Yamashita T, Zhu S. 2017. *On the foundations of ex post incentive compatible mechanisms.* Unpublished manuscript, Toulouse School of Economics, Toulouse, France