Robustness in Mechanism Design and Contracting

Gabriel Carroll

Stanford University; email: gdc@stanford.edu

Version: April 24, 2018

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–31

https://doi.org/10.1146/((please add article doi))

Keywords

TO BE FILLED

Abstract

This survey summarizes a nascent body of theoretical re
of incentives when the environment is not fully known
and offers some general conclusions from the work so
that such models of uncertainty and robustness should
as a niche topic, nor as a panacea for shortcomings of
but instead simply as an additional set of tools in the to
ment more traditional, fully-Bayesian approaches. The
that such models can offer, and the methodological an
lenges they confront, broadly parallel those of traditio

1. INTRODUCTION

This survey gives an overview of a recently growing body of theoretical research on robust design of incentives when details of the environment are not fully known to the designer. A hallmark of this work, distinguishing it from more traditional mechanism design, is that the designer is assumed to be non-Bayesian — at least some aspect of her uncertainty is not expressed using a probabilistic belief.

The survey is intended both to give a succinct guide to the somewhat dispersed work that currently exists, and to try to draw some general lessons from the contributions and 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, or political economy.

The survey will not assume specific technical background in mechanism design and contract theory, although it will be helpful for motivation to have at least a passing familiarity with the questions studied in this field.

A quick note on terminology: 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. When the word "contract" is used, it will be for reasons of (application-specific) precedent rather than principle.

1.1. Non-Robust Models

To illustrate some of the motivation for robust modeling in mechanism design, it will help to begin with a couple of 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 will produce a stochastic amount of output for the principal. The agent will choose an action a, which consists of either exerting high effort or low effort, $a \in \{H, L\}$.[1] Output y will follow 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 $u(w)-v(a)$. Here $g(\cdot)$ and $u(\cdot)$ are utility functions, and $v(\cdot)$ is a cost-of-effort function, with $v(H) > v(L)$. The principal's problem is to choose the function $w(y)$ optimally, 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.

[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.

The optimal contract in this model 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 done the right thing. An extension of this logic leads to what has become called the "informativeness principle": if there are any other observable outcomes besides y whose distribution depends on a, even if they are not payoff-relevant for the principal, an optimal contract should condition on them.

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. Nor, probably, would we recommend that someone write a contract based explicitly on formula (1.1). Indeed, we would be hard-pressed to write such a contract even if we wished to, given not only the strained assumption that the agent is limited to two effort choices (or, as in Holmström's formulation, that effort is a one-dimensional variable), 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. This suggests that if we seek to advance the theory to a point where it can speak to the form of real-world incentive contracts, 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 wish to consider such a framework and see whether it can deliver new insights.

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, so we speak of them as being alternative implementations of "the" optimal auction. 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. Note that while we have assumed specific distributions for the buyers' values, the prescription is less sensitive to these distributions than in the moral-hazard model above, in the sense that the distributions matter only through a single parameter (the reserve price).

So far so good. Now consider a variant: the buyers' values are no longer drawn independently. Instead, with probability $1/2$, the two buyers' values are drawn independently uniform on $[0, 1]$ as before; with remaining probability $1/2$, a *single* 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 case, the designer can extract the full first-best as follows: Ask each bidder 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 make 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) (for the case of finite type spaces; and to McAfee & Reny 1992 for continuous types). 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 pervasive 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) presented them as commentaries on modeling methodology. Indeed, arguably the lasting value of these results has been to serve as a guide to subsequent modelers, showing that we usually had better assume independent types or else things can quickly go off the rails.

On the other hand, one might very well think that there are other situations where a designer wants to extract information contained in agents' beliefs, and can indeed do so via side bets; this may be possible if an agent's reported belief is not also used to make allocation decisions that create large countervailing incentives. This idea is the basis of an extensive literature known variously as "Bayesian truth serum" or "peer prediction" (Prelec 2004, Miller et al. 2005, Witkowski & Parkes 2012).

1.1.3. Discussion. In the preceding examples, the optimal mechanism was sensitive to the details of the environment used as inputs to the model. The notion that realistic mechanisms should not be tailored in this way is often referred to as the "Wilson doctrine" or "Wilson critique" (e.g. Maskin 2003, Satterthwaite & Williams 2002, Perry & Reny 2002, Baliga & Vohra 2003) — somewhat strangely, since the quote from Wilson (1987) usually cited for this attribution says that game-theoretic analysis of any given mechanism should not rely on strong common-knowledge assumptions, but does not say anything about the *design* of a mechanism.[2] Regardless of its origins, this criterion of "detail-freeness" has emerged as a litmus test for the reasonableness of a mechanism. Much of the robustness literature has proceeded by making less stringent assumptions about the inputs to a mechanism design model — in the models above, and many others — and seeing whether simpler, more detail-free mechanisms do indeed appear as a result. As we shall see, this project is sometimes, but not always, successful.

1.2. Perspectives on Mechanism Design

As the examples above illustrate, seemingly non-robust models can be quite instructive for some purposes but unreasonable for others. The usefulness of a model depends on what we wish to get out of it. Accordingly, before diving further into the literature, it will help to overview the various purposes that research in mechanism design can serve.

The 2007 Nobel Prize announcement for this field (Royal Swedish Academy of Sciences 2007) states: "Mechanism design theory asks: ... What resource allocation mechanism produces the best attainable outcome? .... It provides a tool for characterizing the optimal institution for any given set of conditions, thereby enabling a much deeper scientific analysis of the merits of alternative institutions." We can break this down into several possible goals of such analysis:

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 — driven by applications such as medical residency matching systems (or other centralized labor markets) and allocation of students to public schools — has this flavor. A classic example is Roth & Peranson (1999) on mechanisms for matching markets with couples. Much of the mechanism design work in algorithmic game theory also has this flavor (see Nisan et al. 2007).
   A different perspective on the "design" interpretation of mechanism design is that the models used are simply stylized representations of some economic interaction, intended to deliver qualitative insights that can guide the practitioner. 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 application of the formula that

[2] Chung & Ely (2007) explain the distinction in detail. In personal communication with this author, Robert Wilson has expressed that he agrees in spirit with the "Wilson doctrine," but is surprised to receive credit for it.

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. This view is most relevant for those who see economics principally as a social science, trying to comment on the observed world; but it is also indirectly useful for design, insofar as designers often have a choice of what model to write down, and one way to assess the appropriateness of a model is whether it makes predictions that match reality in known situations.

   For example, another workhorse of mechanism design is the Mussa-Rosen model of price discrimination by a monopolist who can produce goods at multiple levels of quality (Mussa & Rosen 1978). The key prediction of this model is that low-value buyers should end up buying inefficiently low-quality goods, i.e. their marginal willingness to pay for a unit of quality is higher than the marginal production cost of quality. The monopolist sets prices that induce this distortion in order to extract more surplus from high-value buyers. Textbook treatments emphasize this as an explanation for observed price discrimination practices (often adorned with a quote from Dupuit in 1849 about miserable conditions in third-class train carriages.)

   This "explaining observations" view of mechanism design can also be applied at the literal level, to explain specific forms of incentives, as well as the qualitative level. An example is Holmström & Milgrom's (1987) model to explain the pervasiveness of linear incentive contracts — that is, why agents are often rewarded by a linear function of the value they produce, rather than by other formulas such as likelihood ratios.
3. A very different view of mechanism design, emphasized 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 bilateral trade model (Myerson & Satterthwaite 1983) is that, when both parties to a transaction hold private information about their own preferences, and neither can be forced to participate in the transaction, 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, for example, as a benchmark to evaluate the performance of actual institutions (e.g. Larsen 2018).
4. Finally, mechanism design can provide simple modeling tools that can serve as an input for studies of other economic phenomena. For example, the literature on organizational form surveyed in Mookherjee (2006) emphasizes agency frictions due to asymmetric information as the cost of decentralized decision-making. To compare the efficiency of 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 incentive 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, and we shall occasionally refer back to this list later. To the extent that actual designers may have limited information about their environment, the relevance to items 1 and 2 on the list is fairly obvious. The relevance to item 3 is less clear-cut; but as we shall see, one connection is that studies of robustness can be

used to ask what kinds of environments are most adversarial, without making assumptions on the mechanism used. Item 4 is also less obvious, but one benefit of the robustness literature is that it sometimes provides new and tractable models to study problems that were difficult under traditional Bayesian models, or for which no canonical Bayesian model existed. This property may make such robust models useful as inputs into other models, regardless of whether the modeler is interested in uncertainty and robustness per se.

1.3. Overview

The next section describes in detail various directions in which robustness has been studied in mechanism design.

The problems studied in mechanism design are quite diverse. The bulk of the literature considers settings of *private values*, in which agents know their own preferences; but some work considers *interdependent values*, where an agent's preferences can depend on information held by other agents. Some work studies *implementation* questions, in which a designer has some target outcome (or a set of acceptable outcomes) for each possible profile of agents' types, and wishes for a mechanism that ensures a desirable outcome; other cases are concerned with maximizing some numerical objective, such as expected revenue or welfare. The discussion below is organized loosely based on the dimensions along which robustness is desired, and largely ignores these other divisions, hopping back and forth across them as convenient.

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 are approximately optimal across many environments, without seeking any kind of exact optimality. Hartline (2012) gives a succinct argument for why such studies can deliver important qualitative insights, just as exact-optimality results can. This literature will not be covered here (except for some comments on mechanisms that attain exact optimality in the large-population limit); the reader is instead referred to the survey by Roughgarden and Talgam-Cohen (roughtcsurvey) in this issue.

One commonly-taken approach to robustness has been to directly restrict the allowable mechanisms in some way that is clearly sufficient to ensure the desired robustness, then ask what can be done under the assumed restriction. By far the most widespread example of this (though not the only one) is the use of dominant-strategy mechanisms to ensure robustness to uncertainty about agents' beliefs, as will be discussed more in Subsection 2.2. This survey will largely avoid this approach, instead emphasizing work that models the desire for robustness more explicitly (including asking whether such foundations can actually justify the restrictions that have been imposed). Similarly, the survey will mostly steer clear of interesting work (such as Chassang & Padro i Miquel 2016 and Chassang 2013) that provides prior-free properties of particular mechanisms.

2. ROBUSTNESS IN MECHANISM DESIGN

2.1. Robustness to Technology or Preferences

The model of Carroll (2015), which studies a variant of the moral hazard problem from Subsection 1.1.1, exemplifies the robust approach. First, 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, but the contract remains sensitive to assumed distributions; for example, with just two actions $\{H, L\}$, the optimum puts all payment on the output with the highest likelihood ratio, and pays zero for any other output.

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 will result 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. (Here an action is represented by its cost to the agent — $v(a)$ in the earlier model — and the resulting distribution over output.) 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 y. Thus, linear contracts provide the most robust way of aligning the agent's interests with the principal's.[3]

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: namely, given a contract w, the agent will never choose a distribution that pays him less on average than the payoff he could get by choosing among the *known* actions, since such a distribution would certainly be suboptimal. That is, the principal knows the F chosen by the agent will satisfy a lower bound on $\mathbb{E}_F[w(y)]$. 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.

How does one interpret the maxmin criterion? One could take it literally, as a description of a principal's decision-making. The interpretation preferred in Carroll (2015) is instead 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 $\mathcal{A}$). The fact that linear contracts not only give some such guarantee, but actually give the *optimal* guarantee, gives a sense in which they are distinguished. This property may help explain why linear contracts are widespread in practice, even if nobody is explicitly optimizing a maxmin objective.

[3] A number of other works, most prominently Holmström & Milgrom (1987) and Diamond (1998), have given foundations for linear contracts in Bayesian models, following somewhat different intuitions than the one here.

The machinery and result of the above model can readily be carried to more complex models. Dai & Toikka (2017) consider a problem where a principal writes robust contracts for a team of agents, where each agent $i = 1, \ldots, n$ will 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, 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 will be 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 will produce 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. Jin (2018) considers robust relational contracting between a long-lived principal and a series of agents, where the principal can propose a payment rule each period but may renege ex post, and shows that optimal payment rules take the form $w(y) = \min\{\alpha y, b\}$, where the bound b is determined endogenously by the no-reneging constraint. Carroll (2017a) considers a setting of information acquisition: rather than producing output directly, the agent acquires costly information about an underlying state of nature, which the principal will use to make some investment decision, and the agent can later be paid based on how well his information matches the realized state. The principal has only partial knowledge of the agent's options for acquiring information. There is a natural way to define linear contracts here: the agent acquires information, recommends an optimal investment, and then is paid a fraction α of the ensuing returns. It turns out that maxmin-optimal contracts are a somewhat subtle variant on such linear contracts.

There has been other work considering alternative ways in which 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 in output, with unknown coefficient. In this case, no guarantees on the principal's profit are possible, since when the agent's cost coefficient becomes unboundedly large, the total surplus available becomes arbitrarily small. So instead they considered guarantees on the *ratio* of the principal's payoff under the chosen contract to the payoff that she would have gotten if she knew the true environment (i.e. the agent's cost function). They showed that the best 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 can exert effort to reduce the costs spent, and also has private information on his "intrinsic" cost (the cost achieved by exerting no effort). The government can offer a contract that specifies payment as a function of reported cost. By making the function flatter, the government can give the supplier more incentive to exert effort (and pocket the resulting savings), but may be leaving unnecessary rents on the table if the intrinsic cost is low. Laffont and Tirole's analysis showed that it is generally optimal to offer a menu of 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 as in the original model, but has maxmin-style uncertainty about the effort cost function. The only substantive assumption about this function is that there is some lower bound $\underline{k}$ on the net efficiency gain (i.e. cost savings minus effort cost) that the supplier can generate. Garrett shows that the maxmin-optimal menu now consists of just two contracts, one that pays a fixed price and one that reimburses costs one-for-one, as opposed to the more complex menus in Laffont and Tirole's original model.

Frankel (2014) applies uncertainty about the functional form of agent preferences in a model of multiple delegated decisions. An agent prepares to face N similarly-structured decisions; in each decision i, she receives private information θ_i and takes an observable action a_i. This will result 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.).

2.2. Robustness to Beliefs and Strategic Behavior

Probably the most extensively-studied form of robustness in mechanism design is robustness to agents' 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 the approach has been applied in many other places, 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 happens without them.

The usual alternative methodology is to design *dominant-strategy* (or *strategy-proof*) mechanisms, i.e. ones in which agents are asked to report their preferences, and each agent's optimal strategy is to report truthfully no matter what other agents report. Thus, the mechanism can be analyzed without any assumptions on beliefs. There is a long tradition in social choice that simply takes as axiomatic that mechanisms should satisfy this property — going back to Gibbard and Satterthwaite's impossibility theorem (Gibbard 1973, Satterthwaite 1975), which essentially says that the only dominant-strategy voting mechanisms are dictatorships. The dominant-strategy approach is natural in social choice domains such as voting or matching, where the space of preferences is unstructured enough that there is no obvious way to formulate a prior; and has recently gained resurgence with the popularity of 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. This work will not be surveyed here; see for example Barberà (2001) or Sprumont (1995) for an overview. Instead, the focus here will be on the newer work that studies whether robustness to beliefs — or to related forms of uncertainty about what agents expect each other to do — actually requires using the dominant-strategy property. That is,

rather than directly imposing the dominant-strategy property in order to ensure robustness, this literature formally models robustness as the objective and asks how best to achieve it.

To see why this can make a difference, consider the following example based on Bergemann & Morris (2005).[4] 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:

	a	b	c	a'	b'	c'
θ_1	2	-1	0	-1	2	0
θ_1'	0	0	1	0	0	1
θ_2	0	2	1	0	1	0
θ_2'	1	0	0	2	0	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 choose 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* 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 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 where voters have cardinal utilities over outcomes, and voting mechanisms may be randomized. Hylland (1980) first proved a version of the Gibbard-Satterthwaite impossibility theorem for such environments, showing that in effect the only dominant-strategy mechanisms are random dictatorships (a voter is chosen at random according to an exogenously fixed distribution, and gets to pick his favorite outcome). 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 enforce a return to random dictatorship. They observe that equilibrium play of this mechanism weakly dominates pure random dictatorship,

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

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 impossibility result. On the other hand, if one requires improvement over random dictatorship at the ex post stage (i.e. for every profile of all voters' preferences and beliefs) rather than the interim stage, they show that such improvement is no longer possible.

This leaves open the question of whether the gap between belief-robustness and dominant-strategy mechanisms arises in more classical models where a designer optimizes some numerical objective, such as revenue. 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. (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. This construction has later been generalized by Chen & Li (2017) and Yamashita & Zhu (2017).) Chung & Ely (2007) also give a counterexample to show that without their regularity assumption, their main result can fail: for any hypothesis that the seller may entertain about the buyers' beliefs, she can construct a side-betting mechanism that does strictly better than the best dominant-strategy mechanism.

Carroll (2018) considers a planner who not only is unsure about the beliefs that mechanism participants have about each other, but is also concerned about resource costs spent in reaching those beliefs — either acquiring information, or influencing the information of others. Carroll considers a planner designing a mechanism for trade, with a prior over agents' values, whose objective is worst-case expected welfare; here welfare reflects both surplus in the trading mechanism *and* costs spent manipulating information beforehand, and the worst case is over possible games by which such information can arise. Note that dominant-strategy mechanisms give no incentive to manipulate information in any way, since doing so cannot change play in the mechanism. In a simplified bilateral trade environment, the maxmin mechanism can be fully characterized: for some parameters, it is a dominant-strategy mechanism; but for others it is a non-dominant-strategy mechanism where one party chooses a take-it-or-leave-it price offer to the other.

Brooks & Du (2018) consider robustness to beliefs in (pure) common-value auctions, building on earlier work by Du (2018) and Bergemann, Brooks & Morris (2016). That is, unlike the private-value environments above, they assume the value of the good to all buyers is the same, but each buyer does not know this value, and instead only observes some noisy information about it. Unlike the private-values setting, there is no canonical model of such environments in the literature. Brooks and Du study the problem of an auctioneer who wants to maximize worst-case expected revenue, where the expectation is with respect to a (fixed) prior over the good's value, and the worst case is over information structures, describing what the buyers know. (A technical difference from the works above is that Brooks and Du assume the buyers share a common prior.) In a technically deep

analysis, they identify the optimal mechanism and the worst-case information structure. The information structure can be summarized as follows: each agent receives a signal that is a positive real number; these signals follow exponential distributions, independently across agents; and the good's value is correlated with the signals in such a way that it depends only on the sum of the signals. The optimal mechanism is harder to describe, and one would not take the result as a literal prescription for auction design. But one view of their work — 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.

Carroll (2016) considers a related problem of uncertainty about information structures in a simpler, binary-outcome environment: two players can either accept or reject a mutual agreement, whose net payoff to each is uncertain and may be negative. There is the possibility for adverse selection: one player accepting the agreement can be a signal that it is bad for the other player. Carroll (2016) identifies a robust lower bound on the equilibrium probability of successful agreement in spite of adverse selection. Although the focus is on analysis of a fixed mechanism, one consequence is that this simple accept/reject mechanism is actually maxmin-optimal in some cases.

The works above take a "structural uncertainty" approach, which follows the orthodoxy of assuming equilibrium behavior, and models the designer's uncertainty as uncertainty 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. (Yamashita 2015b offers a link between the two approaches; see also Bergemann & Morris 2011.)

Börgers (1991) first showed that such an approach can be more permissive than the dominant-strategy approach, and in intuitively plausible ways. Namely, consider a voting setting: there are several possible outcomes, and each agent (voter) may have any arbitrary preference ordering over the outcomes. However, Börgers points out that with three possible outcomes, *approval voting* (with an appropriate provision for tie-breaking) guarantees a Pareto-efficient outcome whenever the voters play undominated strategies, and treats the voters fairly, thus overcoming the pessimistic conclusion of the Gibbard-Satterthwaite impossibility theorem. (In approval voting, each voter can approve any subset of outcomes, and whichever outcome gets the most votes wins. A voter's undominated strategies are either to approve his most-preferred outcome or his two most-preferred outcomes.)

There are other works exploring both the possibilities and limits of undominated-strategy implementation. Babaioff, Lavi & Pavlov (2006) offer a mechanism for certain combinatorial auction settings, that provides a nontrivial welfare guarantee in undominated strategies, where no nontrivial dominant-strategy mechanism is known. Jackson (1992) shows that for any *bounded* mechanism (a well-behavedness restriction), the outcomes it implements in undominated strategies must satisfy a weak condition termed *strategy-resistance*, which specializes to coincide with strategy-proofness if the mechanism always ensures a unique outcome. Börgers & Smith (2012), in analysis 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.

A particular problem that has been much studied is that of implementing a *social choice function*, where for each type profile of the agents, the planner has a unique outcome that

she wishes to ensure. (More background on this problem, including an example, will appear in Subsection 2.5.) Which such functions can be implemented? If agents' preferences are privately known and can vary independently, Jackson (1992) implies that the desired outcome can be ensured for all undominated strategies only if it can be implemented in dominant strategies. One can alternatively consider settings where agents have complete information about each other's preferences. The complete information assumption has force if we are willing to assume agents apply iterated deletion of dominated strategies (i.e. rationalizability), rather than just one round of deletion. Abreu & Matsushima (1992) show that in many environments, any social choice function can then be *virtually* implemented, meaning that we can ensure the desired outcome with probability $1-\epsilon$ for arbitrarily small ϵ. However, their mechanism has been criticized for unreasonable reliance on many rounds of deletion (Glazer & Rosenthal 1992). (See also Abreu & Matsushima (1994), Bergemann, Morris & Tercieux (2011) for exact implementation.)

If information is not complete but preferences are interdependent, then the question of which social choice functions can be implemented under iterated deletion again becomes nontrivial. Bergemann & Morris (2009b) give a general answer for virtual implementation, and Bergemann & Morris (2011) for exact implementation. However, the mechanisms constructed there appear rather esoteric. In more specific environments, cleaner results are possible: Bergemann & Morris (2009a) consider a model in which each agent's preferences depend only on a one-dimensional aggregate of all agents' types, and the preferences satisfy a single-crossing property. They show that implementation in rationalizable strategies is possible if and only if it can be done using the direct mechanism where agents just report their type. The relevant necessary and sufficient condition is a "contraction" property that essentially limits the amount of interdependence.

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 as in Chung & Ely 2007). 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. In the process, Yamashita shows how much of the classical technical machinery from auction theory can be adapted to study undominated-strategy mechanism design.

When there are mechanisms that can do robustly better than dominant-strategy mechanisms, it is usually not known what the optimal mechanism is. As an alternative to 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 mechanisms. 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.

On the opposite side, some argue that even dominant-strategy mechanisms are not enough to ensure robustness in practice, because human players may fail to play their dominant strategy. Li (2017) argues that mechanisms should be *obviously strategy-proof*:

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 by deviating. Thus, each agent can see that his strategy is optimal without needing to perform the contingent-reasoning exercise of imagining a deviation while holding others' strategies fixed. For example, 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. This criterion has recently been applied to a number of mechanism design domains; see Pycia & Troyan (2017), Ashlagi & Gonczarowski (2016).

One more approach to accommodating uncertainty about strategic behavior is to consider settings where agents will interact repeatedly using the same mechanism; one might then hope that they will learn about each other's actions and adjust accordingly. Then, one can aim to design mechanisms where learning dynamics will lead to desirable outcomes in the long run. The large literature on learning processes in games has mostly found that one cannot usually expect such processes to converge to equilibrium, except in some specific classes of games, see e.g. Fudenberg & Levine (1998). So, convergence becomes an additional desideratum for the designer. For example, Healy & Mathevet (2012) consider mechanisms in which the best-reply mapping is a contraction. Perhaps the closest work to practical applications here is that of Sandholm (2002, 2005, 2007), arguing that in Pigouvian-style congestion pricing mechanisms, various natural dynamics will converge to socially efficient equilibrium outcomes.

2.3. Robustness to Distributions

Design problems commonly involve maximizing the expectation of profit (or some other objective) with respect to a known prior distribution over agents' preference types. It can be natural to ask what happens if the designer has only partial information about the distribution, and so wishes to maximize a guarantee under this partial 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 the previous subsection, 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 satisfying these constraints? Intuitively, the seller would want to randomize the price in order to hedge the uncertainty. This is in contrast to the case of a known distribution, where it is always optimal to set a single, deterministic price (Riley & Zeckhauser 1983). Randomizing over prices is equivalent to offering a menu of probabilities, i.e. screening the buyer types by allowing the buyer to pick a probability q of receiving the good, and specifying a price $p(q)$ for each such probability (with $p(q)$ a nonlinear function of q). Carrasco, Luz, Kos, Messner, Monteiro & Moreira (2017a) explicitly derive the optimal distribution over prices. (They also consider a generalization to where multiple moments of the distribution are known, although in this case the parameters of the optimal mechanism cannot be given explicitly.)

Many variants of this problem quickly present themselves. Another paper by a subset

of these authors (Carrasco et al. 2017b) considers a version where the monopolist can sell continuous quantities and the agent has nonlinear preferences, and characterizes 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 monopoly 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.

In models of this sort, the question arises as to exactly what one learns from the exercise. The idea of hedging uncertainty by randomizing is natural (and familiar from the theory of zero-sum games); and otherwise, qualitative properties (such as distortion below first-best) are often similar to those from the Bayesian model. Ideally, one would like to be able to give some economic interpretation to the specific form of the optimal mechanism.

One attempt to study such a problem with a more concrete interpretation is Carroll & Meng (2016b). They study a principal-agent problem, and pursue a possible robustness justification for linear contracts, based on the idea that such contracts give the same incremental incentive for effort at every point along the contract. In their model, the principal contracts on output, which is the sum of (one-dimensional) effort and a random noise term; but the agent chooses effort *after* observing the noise. This is thus a "false moral hazard" model, which can be subsumed in the framework of Carrasco, Luz, Monteiro & Moreira (2017b) (and 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 to satisfy a limited liability constraint.

One 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 a menu of infinitely many probabilistic bundles at different prices (Daskalakis et al. 2013). Carroll (2017b) 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 present a simpler, duality-based proof for the monopolist problem, and also give an extension to budget-constrained buyer.)

Dworczak (2017) uses a distributional-robustness argument to motivate a particular class of mechanisms. He studies a setting of mechanism design for agents who will later participate in some further interaction that depends on information revealed by the mechanism. Characterizing optimal mechanisms is challenging, but he restricts to a tractable class of *cutoff* mechanisms, which have the property that the associated allocation rule 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 kind of guaranteed implementability may be desirable.

In settings where the distribution of agent types is unknown, but samples are available — by looking at previous (or even simultaneous) participants in the mechanism — the designer would naturally want to use a mechanism that learns the distribution from these samples. (Indeed, this is the case also in a Bayesian model, with a prior distribution over distributions. However, in practice, specifying such a prior can be unwieldy except for specific parameterized cases.) A number of "folk mechanisms" formalize versions of this idea. For example, in a market with a large number of agents seeking to exchange goods, one can split the market into two (or more) submarkets, ask agents to report their preferences, and have trading prices in one submarket determined by Walrasian equilibrium in the other; agents have no incentive to misreport their preferences because they cannot change their own prices. Various mechanisms of this form can be shown to converge to efficiency as the number of participants becomes large; see e.g. Kovalenkov (2002), Hashimoto (2016) and Kojima & Yamashita (2017). Segal (2003) studies a seller selling identical goods with a nonlinear production cost, derives the exactly-optimal mechanism when the designer has a Bayesian prior over distributions, and compares its rate of convergence to optimality against some intuitive prior-free approaches.

For such learning problems in prior-free settings, beyond simply establishing asymptotic optimality, a deeper technical question is how to use the samples efficiently to ensure good rates of convergence, uniformly across a wide class of possible distributions. Finding exact maxmin-optimal mechanisms seems to be intractable, so much of the literature (pioneered by Roughgarden and co-authors, e.g. Cole & Roughgarden 2014, Huang et al. 2015, Morgenstern & Roughgarden 2015 and Roughgarden & Schrijvers 2016) has explored how to achieve near-optimal convergence rates. This involves technical constructions to efficiently hedge against the possibility of drawing samples that are unrepresentative of the true distribution.

Of course, another approach to learn distributions is simply to elicit information about the distribution from the agents. If the agents' beliefs are assumed to come from a common prior, then in principle one could just ask the agents to report the prior, punish them all (say by canceling the mechanism) if the reports disagree, and if the reports agree then run the optimal mechanism for that prior. 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 allowing bidders to propose reserve prices for other bidders. Brooks (2013) considers a single-good auction in which buyers' types need not be independently distributed. He considers the class of all joint distributions satisfying a bound on the ratio of the highest possible value to the expected surplus available, and formulates the problem of maximizing the ratio of the expected revenue to expected surplus, in worst case over such distributions. He shows this is achieved by a "surveying and selling" mechanism, in which each buyer is asked 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. A perturbation of this mechanism provides strict incentives and so guarantees that the only rationalizable behavior is truthful reporting (both of values and of beliefs).

2.4. Robustness to Collusion and Renegotiation

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* — that is, to coordinate their reports in the mechanism (perhaps after exchanging some information with each other) so as to achieve a jointly better outcome. Collusion is a major concern in auction practice. And they may also be able to *renegotiate* (or *reallocate*) ex post: to jointly agree to change the outcome specified by the mechanism, if some other outcome is more to their liking. 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: *group strategy-proofness*, which requires that no coalition of agents can ever jointly misreport their preferences in a way that makes each of them better off. This criterion is quite demanding, but in numerous environments there are interesting mechanisms that satisfy it (e.g. Bird 1984, Dubins & Freedman 1981), and it can even be implied by individual strategy-proofness in some situations (see Le Breton & Zaporozhets 2009 and Barberà et al. 2010).

In settings with monetary transfers, such as auctions, one might imagine that members of a collusive coalition would make side payments among themselves, in which case group strategy-proofness is not enough and coalitions would instead 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 his individual preferences and the set of other agents he is colluding with, and propose a single-good auction where truthful reporting is a dominant strategy, in the sense that no coalition can benefit by jointly misreporting. Chen & Micali (2009) achieve nontrivial guarantees in a combinatorial auction environment under a much more agnostic model of coalition behavior.

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 following Laffont & Martimort (1997), they allow a quite general class of collusive protocols, but assume 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 side contract, wherein they exchange information about their types, and the side contract may specify how they will manipulate the mechanism *and* a possible reallocation from the outcome specified by the mechanism; but agents can strategically misrepresent their types in the side contract, 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 that makes the designer's payoff become independent of the realized agent types. 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 consider 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.

Both Che & Kim (2006) and Che & Kim (2009) require implementations that are not ex post individually rational — agents sometimes end off 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 adopt a weaker model of collusion in which bidders cannot make side payments to each other, so that the allocation of payment across individual bidders matters. They characterize the optimal collusion-proof auction within this class, which is generally strictly worse than the seller could do without collusion.

Note that the optimistic results in Che & Kim (2006, 2009) rely on shared beliefs among the designer and all agents, and so are subject to the usual criticisms of making strong assumptions about beliefs. If one added a concern about collusion to a belief-robust model as in Chung & Ely (2007), the collusion constraint would be binding.[5]

In the study of mechanism design with renegotiation, the 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, due to each agent's ability to 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.

In practice, 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 the dominant-strategy property 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. While their approach provides a strong guarantee

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

of robustness against renegotiation, it is potentially open to the same critique as raised for dominant strategies above: one might be able to do better by allowing mechanisms in which agents will behave differently depending on the renegotiation procedure (and may sometimes renegotiate along the equilibrium path).

The one work that has addressed this possibility head-on is Carroll & Segal (2018). They study a single-good auction problem, where a designer 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 will affect bidders' behavior in the auction. 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.[6] Even if there is resale, modeled by including some particular resale game after the auction, typically the optimal auction still involves some discrimination in the auction, 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 bidder (who then does not resell). Thus, contrary to the approach of Neeman & Pavlov (2013), they do not impose renegotiation-proofness a priori, but it emerges in the solution to their maxmin problem.

2.5. 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 will still perform well if the environment is slightly misspecified. In many cases, qualitative properties of the optimal mechanism will be unchanged; the mechanism will look 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 will typically be 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

[6] Che and Kim's (2006) collusion-proofness construction does not apply here: they consider only collusive side-contracts that players agree to before the mechanism, with no information about each other; whereas in post-auction resale, the players necessarily *have* received information about each other, by observing the outcome of the auction.

value $2-\epsilon$, this type will not buy and profit falls to 2/3.) They show that a simple fix — rebating a small fraction of revenue to the buyer — will make 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 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.

Both of the above works studied settings in which a principal interacts with one agent at a time. It appears to be an open question whether similar local fixes apply in Bayesian settings, where one might worry that a slight misspecification of one agent's type distribution could affect other agents' incentives in a destabilizing way.

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 environment. 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 in Subsection 1.2.

One area of mechanism design where local robustness can make a drastic difference is in implementation under complete information (hinted at briefly in Subsection 2.2 above). The most standard formulation of this problem, credited to Maskin (1999), studies *social choice correspondences* and uses Nash equilibrium as the solution concept. The designer has in mind a correspondence specifying one or more acceptable outcomes for each possible profile of preferences, and wishes to design a mechanism for which Nash equilibrium play will result in an acceptable outcome. One option is to ask everyone to report the full state of nature, and punish them all if they disagree; truthful reporting is then an equilibrium, but there are plenty of other equilibria in which the agents coordinate on some false report. So this literature takes up a more demanding goal: to design mechanisms in which *every* equilibrium produces desirable outcomes.

For a concrete example, imagine a buyer and seller who contract to trade some good that has not yet been produced. After it is produced, it will be 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 even though they will both agree on the quality, the court that will enforce the contract lacks the expertise to verify it. They also are not satisfied to use the simple mechanism above, where both parties report the quality and trade at the corresponding

price if their reports agree (and don't trade if they disagree), because the seller may be worried that she cannot escape from the bad equilibrium wherein both parties just always report low quality regardless of the true state. More generally, *no* static mechanism for this problem is free from such bad equilibria: the desired outcome fails a key necessary property known as "Maskin monotonicity" (from Maskin 1999).

Moore & Repullo (1988) proposed a solution, applicable for this example and in general: allow dynamic mechanisms, and assume that agents will play a subgame-perfect equilibrium. They show that this allows almost any outcome to be implemented. Their construction essentially gets rid of false reports by allowing each agent to "challenge" a report by another. In the above example, the construction could operate 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. In equilibrium, the buyer will be dissuaded from trying to get a cheap price by reporting low quality when the true quality is high, because the seller will challenge in order to (successfully) sell at the higher price.

Aghion, Fudenberg, Holden, Kunimoto & Tercieux (2012) note that this kind of mechanism is very sensitive to the assumption of complete information. Suppose each player has just a slight probability of misperceiving the quality. Suppose the buyer still is expected to report (his perception of) the quality truthfully. If the seller thinks the quality is high, but 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. More generally, the key insight is that when a small amount of incomplete information is introduced, reaching a subgame that was previously off-equilibrium-path is now a very informative event, and this can discontinuously change predicted behavior. As shown in Aghion, Fudenberg, Holden, Kunimoto & Tercieux (2012), if we require outcomes to be robust to an arbitrarily small amount of incomplete information, no mechanism can implement an outcome that violates Maskin monotonicity (as in our buyer-seller 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 all equilibria, should give the desired outcome). In the buyer-seller example above, this would allow for the simpler mechanism, asking both parties to report and forbidding trade if they disagree. Oury and Tercieux note that if we allow for local "email-game-style" (Rubinstein 1989) perturbations of beliefs, then even these equilibria can disappear. They show that requiring robustness to such perturbations limits the implementable social choice functions to those satisfying Maskin monotonicity.

2.6. Robustness of Standard Mechanisms

There is also a body of literature that could be classified under the heading of "robustness," that does not consider design questions, but rather studies ways in which "standard" mechanisms perform well across a range of environments, often approaching efficiency when there are many agents. For example, for a market with a large number of buyers and sellers trading units of a single good, a standard mechanism is the double auction: each agent names a price at which he is willing to buy or sell, and trades get executed at the market-

clearing price (in finite markets there is an interval of market-clearing prices; one can, say, pick the midpoint). Individual agents have incentives to misreport their value to get a better price. But Rustichini, Satterthwaite & Williams (1994) and Satterthwaite & Williams (2002) consider the Bayesian equilibrium of such a mechanism, and show that it converges to efficiency as the market grows large — and moreover, converges at the same rate as the optimal mechanism. Subsequent work has generalized to interdependent values (relevant, for example, for trading in financial assets) or correlated private values; see Cripps & Swinkels (2006) and Reny & Perry (2006). These results are robust in the sense that they require minimal assumptions on the distribution of agents' types.

More generally, given the prominence of Walrasian equilibrium as a canonical mechanism for exchange economies with many goods, there has been interest in studying its efficiency properties when agents behave strategically. Jackson & Manelli (1997) showed that under regularity assumptions, the outcome of strategic manipulation with many agents will be asymptotically efficient, as long as agents' strategic behavior conforms to a kind of "self-confirming equilibrium" assumption — essentially, that each agent be able to justify his manipulation by some (not necessarily correct) belief about other agents' behavior that is consistent with the observed prices.

In contrast with the above results on asymptotic optimality, a more recent literature, much of it in computer science, has studied situations in which standard mechanisms approximate optimal performance to within some modest constant factor across broad classes of environments; see Roughgarden & Talgam-Cohen (2018).

3. DISCUSSION

This is a natural spot to give some reflections — to try to extract some general lessons from the small but rather diffuse body of work so far on robustness in mechanism design, and to comment on what might be the most productive directions for future work, both for individual contributions and for the progress of the field as a whole. These comments will necessarily be more subjective and speculative than the summary above.

3.1. Lessons Learned

Evidently, robust models of mechanism design can serve a variety of conceptual purposes, just as with traditional 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, or otherwise unreasonable (as in the motivating examples from the introduction); 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 (this is the case, for instance, for the information acquisition model of Carroll 2017a); 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.

All this said, one should not assume that when a Bayesian model is unsatisfactory in some way, writing down a maxmin model (or some other form of robust model) will automatically do better. Of course, it is hard to assess how often a modeling approach

succeeds just by looking at the published literature, since one largely sees only the success stories. But in this author's experience, a robust model is no more likely to provide tractable, interpretable, and intuitively plausible solutions than a Bayesian model is. As is usually the case in economic modeling, one often has a choice of many models to describe a given situation, and figuring out which models (if any) will deliver insights is largely a matter of trial and error. Robust models simply provide an additional set of tools to try out.

Such models naturally raise a set of interpretive questions. Not long ago, mainstream economic modeling followed a fairly strict orthodoxy of expected-utility maximization — often justified by some mumbling about the Savage axioms. These days the culture is becoming more pluralistic, partly due to the success of behavioral economics in explaining important phenomena, and partly due to a more general shift away from literalism in the interpretation of game-theoretic models. Opponents of maxmin models (say) might argue that real-life decision-makers usually do not maximize extreme worst-case objectives. But it is also rare to 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 Carroll 2015, summarized in Subsection 2.1.)

That said, robust models do face some extra hurdles relative to Bayesian models, particularly if they are to be incorporated into larger game-theoretic settings (as in item 4 from the taxonomy). For one, if we wish to write equilibrium models in which multiple interacting players face maxmin-style uncertainty (such as the multiple-principals contracting model of Marku & Ocampo Díaz 2017), this demands a somewhat more literal interpretation of the maxmin objective than in a single design problem in isolation. Nash equilibrium implicitly presumes that each player is certain of what the other player will do, yet has non-Bayesian uncertainty about some other aspect of the environment — a combination of assumptions that may strain our credulity more than either assumption alone. 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.

3.2. Effective Robust Modeling

Many of the studies surveyed above can be cast as instances of the following general template.

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.

Again, this recipe will not always succeed. Nonetheless, it seems to be a productive approach for generating insights about why, and in what ways, some mechanisms are robust to uncertainty. In interesting cases, typically, the maxmin criterion will be essential to the analysis, in the sense that one would not simply get the same results by guessing the worst case in advance and solving the corresponding Bayesian problem: either the worst case is hard to guess (for example, in the belief robustness models of Chung & Ely 2007 or Brooks & Du 2018); or there is no unambiguously-defined worst case (technically, the maxmin problem does not have a saddle point, as in the contracting model of Carroll 2015); or the worst case may be intuitive, but still requires work to complete the story, either to analyze the resulting Bayesian problem or to argue for why it is indeed the worst case (as in the auctions-with-resale model of Carroll & Segal 2018).

The recipe does invite the obvious criticism of producing models that are reverse-engineered. But once again, this is not specific to the robust approach. Economic models are formal devices for expressing, and evaluating, ideas about the world, and modeling choices are often tailored to best express whatever ideas the modeler has in mind. The robust approach to mechanism design simply provides one additional set of such tools and such ideas.

As in other kinds of economic models, there is some 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 some (such as Bergemann & Schlag 2008, Chassang 2013, Hurwicz & Shapiro 1978) instead looked at minimizing regret relative to some benchmark. Regret can be measured as a difference, or as a ratio. (In computer science there is a strong tradition of using worst-case ratio measures; this tradition comes from analysis of algorithms, where typically one can only make asymptotic statements, so comparing levels is not meaningful. It is not clear that any parallel explanation applies for mechanism design.) There is also some flexibility in the choice of benchmark. 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.[7]

This recipe, then, provides an easy route for the mechanism design theorist interested in this area to sit down and get to work. And advancement of the theory will depend not only on pure modeling, but also on input from applications, which provide examples of practical incentive problems needing to be solved, candidate solutions or incentive mechanisms actually observed in reality, and perhaps intuitions about why they are used. The job of theory is then to elucidate, explain, and perhaps to improve on these mechanisms.

[7]Börgers (2017) critiques the use of a maxmin criterion to justify particular mechanisms. His argument is that these mechanisms may be weakly dominated: one can attach bells and whistles to make them perform better in some non-worst-case environments; thus the simpler mechanism isn't really selected as optimal. (Börgers & Smith 2012, 2014 are closely related.) A reply is that one nonetheless obtains useful insights from analyzing the maxmin problem and showing that the simpler mechanism solves it, that are orthogonal to the insights obtained from the refined criterion.

3.3. The Future of Robustness in Mechanism Design

What will be the lasting lessons from this body of work on robustness? In what directions can and should the field develop to maximize its impact on economics, or even on incentive design in practice? It is always difficult to speculate about the future impact of foundational research. But with that disclaimer, here are a few thoughts.

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 if the field contributes a few 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 — especially if they can provide tractability for studying problems that are difficult to solve using standard Bayesian models. Indeed, for better or for worse, modeling tools that are easy to use may end up being widely applied regardless of whether the assumptions that underlie them are appropriate to the application.

At present, the state of the literature is rather diffuse and it seems hard to identify such a core set of tools. A more focused end state seems desirable. But this is not necessarily a problem in need of intervention: as the field grows organically and matures, it will likely become apparent which tools can be used repeatedly.

Another possible good outcome 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 present state of the field 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) find some improvements over existing mechanisms, yet without identifying new *optimal* mechanisms or crisp lessons about what mechanisms should be used and why (as in the negative examples from Chung & Ely 2007 and Börgers & Smith 2012).

Outcome (a) is, again, not necessarily a problem at an early stage. One view is that the current state of the field is one of developing modeling tools, and tools should be beta-tested by seeing whether they deliver outcomes that 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 modeling robustness concerns can make a difference. But for the tools to gain widespread use, we eventually need to go beyond scattered counterexamples. For some of the purposes that mechanism design serves (items 3 and 4 in the opening taxonomy, and to some extent 2), optimizing is essential. For actual design (item 1), it is less necessary — and in problems approaching real-world levels of complexity, it may be generally hopeless. But the goal of optimization plays an important role in theoretical modeling, by lending systematization and discipline to the exploration of any given design problem. It seems that this work on robustness will have its best shot at influencing practical design if it can develop models in which optimization is possible and leads to discovery of new mechanisms, or if it can at least suggest principled ways of approaching design 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 with (listed in random order) Stephen Morris, Shengwu Li, Alex Wolitzky, Dirk Bergemann, and Ben Brooks. Weixin Chen provided valuable 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. 1992. Virtual implementation in iteratively undominated strategies: Complete information. *Econometrica* 60(5):993–1008

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

Ashlagi I, Gonczarowski YA. 2016. Stable matching mechanisms are not obviously strategy-proof. Unpublished manuscript, Stanford Univ., Stanford, CA

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

Babaioff M, Lavi R, Pavlov E. 2006. Single-value combinatorial auctions and implementation in undominated strategies, In *SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete Algorithms,* pp. 1054–1063. Philadelphia, PA: SIAM

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

Barberà S. 2001. An introduction to strategy-proof social choice functions. *Soc. Choice Welf.* 18(4):619–653

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

Barberà S, Berga D, Moreno B. 2016. Group strategy-proofness in private good economies. *Am. Econ. Rev.* 106(4):1073–1099

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

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

Bergemann D, Morris S. 2007. An ascending auction for interdependent values: Uniqueness and robustness to strategic uncertainty. *Am. Econ. Rev.* 97(2):125–130

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

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. 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. 2017a. Optimal selling mechanisms under moment conditions. Unpublished manuscript, Bocconi Univ., Milan, Italy

Carrasco V, Luz VF, Monteiro P, Moreira H. 2017b. Robust mechanisms: The curvature case, fgv/epge, rio de janeiro, brazil. Unpublished manuscript,

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. 2017a. Robust incentives for information acquisition. Unpublished manuscript, Stanford Univ., Stanford, CA

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

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

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

Chassang S, Padro i Miquel G. 2016. Corruption, intimidation, and whistleblowing: A theory of inference from unverifiable reports. Unpublished manuscript, New York Univ., New York, NY

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. *Journal of Economic 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

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

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. Unpublished manuscript, Simon Fraser Univ., Burnaby, BC, Canada

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, Univ. Chicago, Chicago, IL

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

Fudenberg D, Levine DK. 1998. *The Theory of Learning in Games.* Cambridge, MA: MIT Press

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

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

Hashimoto T. 2016. The generalized random priority mechanism with budgets. Unpublished manuscript, Yeshiva Univ., New York, NY

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, 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. *J. Econ. Theory* 77(2):354–376

Jackson MO, Manelli AM. 1997. Approximately competitive equilibria in large finite economies. *Rev. Econ. Stud.* 59(4):757–775

Jin X. 2018. Ph.D. thesis, Stanford University

Kojima F, Yamashita T. 2017. Double auction with interdependent values: Incentives and efficiency. *Theor. Econ.* 12(3):1393–1438

Kovalenkov A. 2002. Simple strategy-proof approximately walrasian mechanisms. *J. Econ. Theory* 103(2):475–487

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

Laffont JJ, Martimort D. 2000. Mechanism design with collusion and correlation. *Econometrica* 68(2):309–342

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

Mailath G, Postlewaite A. 1990. Asymmetric information bargaining problems with many agents. *Rev. Econ. Stud.* 57(3):351–367

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. *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress.* ed. Mathias Dewatripont, Lars Peter Hansen, Stephen J. Turnovsky, pp. 1–24, Cambridge, UK: Cambridge Univ. Press

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

Miller N, Resnick P, Zeckhauser R. 2005. Eliciting honest feedback: The peer prediction method. *Manag. Sci.* 51(9):1359–1373

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

Mussa M, Rosen S. 1978. Monopoly and product quality. *J. Econ. Theory* 18(2):301–317

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, 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

Perry M, Reny PJ. 2002. An efficient auction. *Econometrica* 70(3):1199–1212
Prelec D. 2004. A bayesian truth serum for subjective data. *Science* 306(5695):462–466
Pycia M, Troyan P. 2017. Obvious dominance and random priority. Unpublished manuscript, Univ. of Virginia, Charlottesville, VA
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, Schrijvers O. 2016. Ironing in the dark, In *EC '16: Proceedings of the Seventeenth ACM Conference on Economics and Computation,* pp. 1–18. New York: ACM
Roughgarden T, Talgam-Cohen I. 2018. To be filled
Royal Swedish Academy of Sciences. 2007. The Prize in Economic Sciences 2007 — Popular Information. `http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2007/popular.html`
Rubinstein A. 1989. The electronic mail game: Strategic behavior under 'almost common knowledge'. *Am. Econ. Rev.* 79(3):385–391
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. *Advances in Economic Theory: Fifth World Congress.* ed. Truman F. Bewley, pp. 33–77, Cambridge, UK: Cambridge Univ. Press
Witkowski J, Parkes DC. 2012. A robust bayesian truth serum for small populations, In *Twenty-Sixth AAAI Conference on Artificial Intelligence,* 1492–1498. Palo Alto, CA: AAAI
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. Theory* 159:267–279
Yamashita T. 2017. Revenue guarantees in auctions with a (correlated) common prior and additional information. Unpublished manuscript, Toulouse School of Economics, Toulouse, France
Yamashita T, Zhu S. 2017. On the foundations of ex post incentive compatible mechanisms. Unpublished manuscript, Toulouse School of Economics, Toulouse, France