American Economic Review 2020, 110(5): 1243–1273
https://doi.org/10.1257/aer.20190283

A Model of Complex Contracts[†]

By Alexander M. Jakobsen*

I study a mechanism design problem involving a principal and a single, boundedly rational agent. The agent transitions among belief states by combining current beliefs with up to K pieces of information at a time. By expressing a mechanism as a complex contract—a collection of clauses, each providing limited information about the mechanism—the principal manipulates the agent into believing truthful reporting is optimal. I show that such bounded rationality expands the set of implementable functions and that optimal contracts are robust not only to variation in K, but to several plausible variations on the agent's cognitive procedure. (*JEL* D82, D86)

Traditional approaches to mechanism design theory assume that when a designer selects a mechanism, agents understand the underlying game form: the mapping from strategy profiles to outcomes. Under this assumption, designers need only consider the incentives induced by the underlying game. In this paper, I study a mechanism design problem in which an agent's comprehension of the game form is subject to complexity constraints, distorting the mechanism's incentive properties. Consequently, the designer also considers agents' bounded rationality, and may seek mechanisms robust to (or exploitative of) limited cognitive ability.

My analysis is motivated by the presence of extreme complexity in many real-life institutions and contracts. Tax codes and legal systems, for example, consist of many interacting cases and contingencies, making correct identification of the game form a daunting task. Policies for allocating jobs, promotions, financial aid, or other scarce resources can also seem excessively complex. However, the manner in which complexity influences the design and effectiveness of mechanisms is not well understood, and analysis of these issues involves difficult conceptual challenges. What distinguishes complex mechanisms from simple ones? How might agents go about processing them? Can designers effectively manage the behavior of cognitively

* Department of Economics, University of Calgary (email: alexander.jakobsen@ucalgary.ca). Jeffrey Ely was the coeditor for this article. This paper is based on a chapter of my PhD dissertation, completed at Princeton University in June 2017. Earlier versions were circulated under the title "Implementation via Complex Contracts." I am grateful to Faruk Gul and Wolfgang Pesendorfer for their guidance, and to coeditor Jeff Ely and three anonymous referees for their comments. Thanks also to Dilip Abreu, Roland Bénabou, Sylvain Chassang, Kyle Chauvin, Dimitri Migrow, Stephen Morris, Rob Oxoby, Doron Ravid, Ariel Rubinstein, Kai Steverson, Benjamin Young, and participants at the Canadian Economic Theory Conference (Toronto 2018) for helpful conversations and feedback.

[†] Go to https://doi.org/10.1257/aer.20190283 to visit the article page for additional materials and author disclosure statement.

constrained agents? How, and to what degree, can designers accommodate heterogeneous cognitive procedures or abilities?

The starting point of my model is to distinguish between mechanisms and the manner in which they are framed. I assume the designer commits to a mechanism by announcing a *contract*: a collection of clauses, each providing limited information about the mechanism. Fully rational agents combine all clauses to deduce the true mechanism, but boundedly rational agents need not. Rather, they adhere to a given procedure for processing and combining clauses, arriving only at coarse approximations of the true mechanism. My concept of bounded rationality rests on basic principles of framing and procedural reasoning and, as explained in Section IV, is not bound to the domain of mechanism design: it can be reformulated as a general model of non-Bayesian updating and applied to other settings.

To illustrate the procedure, consider the game Sudoku. In this game, a player is presented with a 9×9 table. Some cells are initially filled with entries from the set $D = \{1, 2, \ldots, 9\}$ and the player must deduce the entries for the remaining cells. The rules are that each digit $d \in D$ must appear exactly once in (i) each row; (ii) each column; and (iii) each of the nine primary 3×3 subsquares. Sudoku puzzles are designed to have unique solutions given their initial configurations.

For a standard rational agent, there is no distinction between the initial configuration, together with the rules of the game, and the unique fully resolved puzzle. To him, the combination of a partially filled table and the list of rules simply forms a compact way of expressing all entries. Not so for most (real) people, who understand both the rules of the game and the initial configuration but may find themselves unable to solve the puzzle.[1]

How might an individual go about solving a Sudoku puzzle? Consider Figure 1. Suppose the player notices entry 6 in positions $(3, 2)$ and $(4, 7)$. Then, rules (i) and (ii) block 6 from appearing again in column 2 or row 4 (panel A of Figure 1). Combined with rule (iii), this implies X (position (6,3)) must be 6. He updates the configuration to reflect this (panel B). Looking at the new configuration, he realizes 6 cannot appear again in columns 2 or 3. Applying rule (iii), he deduces that Y (position (8,1)) must be 6, and once again updates the configuration (panel C). He proceeds in this fashion until the puzzle is solved or he gets "stuck."

If agents reason this way, what distinguishes a hard puzzle from a simple one? I propose that in simple puzzles, the player is able to "chip away" at the problem: he can gradually fill in the cells, one at a time, without ever having to combine many rules at once. Above, the player only had to combine three rules with his initial knowledge to deduce $\mathrm{X} = 6$, and three again to deduce $\mathrm{Y} = 6$ once he updated the configuration. In simple puzzles, proceeding in this fashion eventually yields the solution. Hard puzzles, however, inevitably lead players to a configuration where a large "leap of logic" (the combination of many different rules or pieces of information) is required to make further progress. If he cannot perform the required chain of reasoning, the player will remain stuck at such configurations.

[1] In other words, a rational agent is *logically omniscient*: if a collection of facts is known to the agent, so are all of its logical implications. As Lipman (1999) argues, logically *non*-omniscient agents are sensitive to the way information is framed: two pieces of information differing only in their presentation may not be recognized as logically equivalent.

Panel A. Player deduces $X = 6$

5	8						3	
	1			4				6
	6	7	2		8			
		4	7		3	**6**		
8		5				2		
3	2	X		8	5	1	9	
7			8					9
	5	8					1	4
9						5	6	8

Panel B. … then $Y = 6$

5	8						3	
	1			4				6
	6	7	2		8			
		4	7		3	6		
8		5				2		
3	2	**6**		8	5	1	9	
7			8					9
Y	5	8					1	4
9						5	6	8

Panel C. New configuration

5	8						3	
	1			4				6
	6	7	2		8			
		4	7		3	6		
8		5				2		
3	2	6		8	5	1	9	
7			8					9
6	5	8					1	4
9						5	6	8

FIGURE 1. A POSSIBLE SEQUENCE OF DEDUCTIONS IN SUDOKU

My model captures this intuition by combining imperfect memory with limited computational ability. Specifically, agents transition among a coarse set of *belief states* by combining up to K pieces of information at a time. In the Sudoku example, belief states are represented by configurations (partially filled tables), and the agent transitions to a new state whenever he deduces the entry for another cell. Deductions are performed by combining current beliefs with up to K pieces of information at a time, as illustrated above. Agents continue to process information and perform transitions until the puzzle is solved or they get stuck in a state where transitions to finer states require the combination of more than K pieces of information. Agents with a higher K can perform more complex deductions and, thus, solve more difficult puzzles.

Since he transitions only among coarse belief states, the agent typically does not retain all new facts he has derived while processing information. For example, when updating his configuration to reflect $X = 6$, he "forgets" that 6 has been eliminated from column 2 and row 4. Belief states capture this forgetfulness. Note that both elements of the agent's bounded rationality are essential: if K were unbounded or belief states unrestricted, the agent would solve any puzzle and, thus, be indistinguishable from a fully rational agent.

The mechanism design problem involves a principal (the designer) and a single, boundedly rational agent. The principal seeks to implement a function mapping agent types to outcomes, and the agent's type is private information. Both the agent's preferences and outside option are type-dependent. While somewhat restrictive, this setup accommodates a variety of persuasion, allocation, and conflict resolution problems. For example, outcomes might represent different schools, agent types the attributes of students, and the principal a governing body with a particular goal (objective function) of matching student types to schools. Parents have their own type-dependent preferences and outside options (initial allocations), introducing conflict between the principal and agent. Similar conflicts emerge if, for example, outcomes represent tasks, agent types the (unobservable) characteristics of employees, and the principal a manager responsible for assigning tasks to employees.

To achieve implementation, the principal commits to a mechanism (a function mapping type reports to outcomes) by announcing a set of clauses, each providing limited information about the mechanism. Combined, the clauses form a contract

that pins down a single mechanism. Thus, from the agent's perspective, the clauses form a puzzle and the underlying mechanism its solution. Belief states are represented by correspondences mapping actions (type reports) to sets of outcomes, indicating the agent's beliefs about the possible consequences of different actions. Able to combine up to K clauses at a time, the agent transitions to a new state whenever some outcome is eliminated as a possible consequence of some action. Carefully designed contracts (sets of clauses) guide the agent to belief states where truthful reporting appears to be the safest course of action, as per the maxmin criterion.

Restricting to single-agent settings isolates the role of bounded rationality by ruling out strategic considerations: under any mechanism, the agent's outcome depends only on his own action which, in turn, depends on his beliefs about the mechanism. Since the principal cannot induce strategic incentives for truth-telling, very few functions are implementable under full rationality: under any contract, a rational agent deduces the outcome associated with each action and chooses his most-preferred alternative. Thus, any conflict between the preferences of the agent and the objective of the principal renders the situation hopeless. With boundedly rational agents, however, the set of implementable functions is considerably larger.

After establishing a version of the revelation principle in Theorem 1, Theorem 2 fully characterizes the set of implementable functions and identifies a class of contracts that achieve implementation for all admissible K. In particular, a function is implementable if and only if it is *IR-Dominant*. Fix a type θ and suppose both $\succsim_\theta$ and $\succsim_{\theta'}$ indicate that any outcome dominated by x (according to θ) is also dominated by the outside option for type θ'. Then, IR-Dominance requires the outcome from truthfully reporting type θ to be at least as good as x. Theorem 2 establishes that IR-Dominance is a necessary condition for implementability, as well as a partial converse: there exists an integer $\bar{K} \geq 1$ such that any nontrivial, IR-Dominant function is implementable if and only if $K < \bar{K}$.[2] Implementation of such an f for all $K < \bar{K}$ is achieved by a particular contract, denoted $\mathcal{C}_f$. Since it achieves implementation for all admissible K, $\mathcal{C}_f$ is optimal from the principal's perspective: if $\mathcal{C}_f$ does not implement f, neither does any other contract.

Informally, $\mathcal{C}_f$ is the result of a simple design heuristic: minimize the informativeness of each clause subject to the constraint that each clause makes truthful reporting appear optimal. Consequently, $\mathcal{C}_f$ satisfies many robustness criteria. For example, one may introduce randomness, impatience, or costs and benefits of reasoning (thereby endogenizing K) without severely undermining the effectiveness of the contract. I discuss these (and other) robustness properties in Sections II and III. Section IIC conducts comparative static exercises and shows, for example, that the principal can implement any IR-Dominant function for any ability K by expanding the set of actions (messages) available to the agent. Combined, these results formally establish a robust incentive for designers to introduce excess (but constrained) complexity into contracts.

Before proceeding to the model, a few comments on related literature are in order (I defer most of the discussion to Section IV). This paper is part of the emerging literature on behavioral mechanism design. Most of this literature involves agents

[2] Informally, a nontrivial function f is one that would induce some type to misreport (or not participate) if the agent understood that the true outcome were governed by f.

who understand game forms and mechanisms but exhibit nonstandard choice behavior (reference dependence, present bias, etc.) or limited strategic reasoning (e.g., level-k reasoning in games). In contrast, I focus on how cognitive limitations affect the agent's understanding of the mechanism itself.[3] The cognitive limitation is modeled as a sensitivity to the way mechanisms are framed, and is quite distinct from imperfect strategic reasoning (e.g., level-k) because it only affects the agent's perception of the game form. In this sense, my model is most similar to that of Glazer and Rubinstein (2012)—henceforth, GR—which studies persuasion with boundedly rational agents. There are several important differences between this paper and GR. Most notably, our models of bounded rationality have distinct formal and conceptual underpinnings, capturing different ideas of what it means to be boundedly rational. The mechanism design problems are also different: GR focuses on persuasion, while I consider a general implementation problem with type-dependent preferences and outside options. Consequently, our models yield rather different insights. I elaborate on this, as well as other related literature, in Section IV.

I. Model

A. *Outcomes*, *Types*, *Contracts*

There is a single principal and a single agent. Let Θ denote a finite set of agent *types* and X a finite set of *outcomes*. An agent of type $\theta \in \Theta$ has complete and transitive preferences $\succsim_\theta$ over X and an outside option $\bar{x}_\theta \in X$. Let $u_\theta : X \to \mathbb{R}$ be a utility function representing $\succsim_\theta$ and $\bar{x} \coloneqq (\bar{x}_\theta)_{\theta\in\Theta}$ denote the full profile of outside options.

Given a finite set A of *actions*, a *mechanism* is a function $g : A \to X$. Let G denote the set of all mechanisms. Under mechanism g, an agent who takes action $a \in A$ receives outcome $g(a)$.

A *clause* is a nonempty set C of mechanisms. The interpretation of a clause is that it describes a property of a mechanism. For example, the clause $C = \{g \in G : g(a_3) \in \{x_2, x_7\}\}$ may be represented by the statement "the outcome associated with action a_3 is either x_2 or x_7."

A *contract* is a set $\mathcal{C}$ of clauses such that $\bigcap_{C\in\mathcal{C}} C$ is a singleton; let $g_{\mathcal{C}} \in G$ denote the sole member of this intersection. Much like a real-world contract, $\mathcal{C}$ is a collection of statements (clauses), each describing various contingencies of a mechanism. Formally, each clause $C \in \mathcal{C}$ indicates that $g_{\mathcal{C}} \in C$. The requirement that $\bigcap_{C\in\mathcal{C}} C$ is a singleton ensures the contract is not ambiguous: only one mechanism, $g_{\mathcal{C}}$, satisfies all clauses of $\mathcal{C}$. This is a standard assumption in mechanism design, and ensures the contract is enforceable. Finally, note that contracts are defined as sets (not sequences) of clauses because the agent's cognitive procedure, described below, would not depend on the ordering of clauses even if one were specified.[4]

[3] The general framework of mechanism design can accommodate uncertainty about the rules of the game (via appropriate type spaces), but the literature has generally assumed common knowledge of game forms.

[4] Naturally, there are many different ways of representing a set C in formal or natural language, some of which may be more complicated than others. I do not assume the agent's comprehension of a clause is independent of its presentation, but rather that the principal has sufficient expressive power to convey clauses as separate statements

B. *Timing*

First, the principal announces (and commits to) a contract $\mathcal{C}$ defining some mechanism $g_{\mathcal{C}}$. The agent observes $\mathcal{C}$, processes its clauses and arrives at beliefs in the form of a correspondence from A to X: an approximation to the true mechanism $g_{\mathcal{C}}$. The precise manner in which the agent forms beliefs is described in the next section. Given these beliefs, the agent decides whether to participate in the mechanism. If he does not participate, he consumes his outside option. If he participates and takes action $a \in A$, he receives outcome $g_{\mathcal{C}}(a)$, the outcome actually prescribed by $\mathcal{C}$.

C. *The Agent's Cognitive Process*

The agent has both imperfect memory and limited deductive (computational) ability. Memory is represented by a set of feasible belief states, and computational ability by an integer K indicating how many clauses he can combine at a time. Presented with a contract, the agent transitions among belief states as he processes its clauses, gradually refining his beliefs until further improvement requires the combination of more than K clauses.

Formally, a *belief* is a nonempty-valued correspondence $b : A \rightrightarrows X$. An agent with beliefs b has narrowed the possibilities for $g_{\mathcal{C}}(a)$ down to the set $b(a)$. A belief b may be represented by the set $B^b := \{g \in G \,|\, \forall a, g(a) \in b(a)\}$ of all mechanisms contained in b. Let $\mathcal{B}$ denote the family of all such sets B^b. Each $B \in \mathcal{B}$ is a *belief state*. Clearly, there is a one-to-one mapping between belief correspondences and belief states. Given a belief state B, let b^B denote the associated correspondence.[5]

An integer $K \geq 1$ represents the agent's *deductive* (*computational*) *ability*. The agent can combine up to K clauses at a time in order to transition among belief states, starting from the state $B = G$. The next definitions formalize the process. For any finite set S, let $|S|$ denote the cardinality of S.

DEFINITION 1 (K-Validity): *Let $\mathcal{C}$ be a contract and $K \geq 1$. A transition, denoted $B \xrightarrow{\mathcal{C}'} B'$, consists of an (ordered) pair of states $B, B' \in \mathcal{B}$ and a nonempty subcontract $\mathcal{C}' \subseteq \mathcal{C}$. If $|\mathcal{C}'| \leq K$, and*

$$B \cap \left(\bigcap_{C \in \mathcal{C}'} C \right) \subseteq B', \tag{1}$$

then the transition is K-valid.

The idea of Definition 1 is as follows. In state B, the agent believes $g_{\mathcal{C}} \in B$. If at most K clauses belong to $\mathcal{C}'$, then the agent has sufficient computational ability to combine them, revealing $g_{\mathcal{C}} \in \bigcap_{C \in \mathcal{C}'} C$. Then, by (1), the agent deduces $g_{\mathcal{C}} \in B'$.

in a way the agent understands. As we shall see, "optimal" contracts are robust to the possibility that the agent fails to process some (even most) clauses.

[5] Since beliefs are represented by correspondences, the agent's beliefs indicate the possible outcome(s) associated with each action, but not any "correlations" between the outcomes of different actions. I relax this assumption in Section IIIC.

Thus, the agent is capable of transitioning from state B to B' by processing $\mathcal{C}'$ and combining the result with beliefs B.

DEFINITION 2 (K-Reachability): *Let $\mathcal{C}$ be a contract and $K \geq 1$. A state $B \in \mathcal{B}$ is K-reachable if there is a sequence*

$$G = B^0 \overset{\mathcal{C}^1}{\to} B^1 \overset{\mathcal{C}^2}{\to} B^2 \overset{\mathcal{C}^3}{\to} \cdots \overset{\mathcal{C}^n}{\to} B^n = B$$

of K-valid transitions.

Definition 2, like Definition 1, is a statement about the deductive capabilities of the agent. A state B is K-reachable if an agent with no initial knowledge of $g_{\mathcal{C}}$ can deduce, through a series of K-valid transitions, that $g_{\mathcal{C}} \in B$. Importantly, the deduction is sound: $g_{\mathcal{C}}$ actually belongs to B if B is K-reachable. Thus, in K-reachable states, the agent does not erroneously eliminate the true mechanism from consideration.

DEFINITION 3 (Induced Belief State): *Let $\mathcal{C}$ be a contract and $K \geq 1$. A state $B \in \mathcal{B}$ is an induced belief state if it is K-reachable and there is no K-valid transition $B \overset{\mathcal{C}'}{\to} B'$ such that $B' \subsetneq B$.*

An induced belief state is a state B that is reachable but unrefinable by an agent of ability K: upon reaching state B, there are no K-valid transitions to strictly finer states. Thus, induced belief states are those where the agent gets "stuck." Importantly, induced belief states are unique.

LEMMA 1: *For every contract $\mathcal{C}$ and ability $K \geq 1$, there exists a unique induced belief state.*

Given a contract $\mathcal{C}$ and an integer $K \geq 1$, let $B_{K,\mathcal{C}}$ denote the unique induced belief state. The associated correspondence, denoted $b_{K,\mathcal{C}}$, is the *induced belief*. These are the beliefs the agent forms about the mechanism $g_{\mathcal{C}}$ before deciding whether to participate and, if so, which action to take.

The concept of induced beliefs captures the idea that the agent, starting from state G, repeatedly processes clauses and performs transitions until he is unable to further refine his beliefs. Since induced beliefs are unique, the procedure is path-independent: there is no possibility of getting stuck in a state other than $B_{K,\mathcal{C}}$, and therefore the order in which transitions are performed does not matter. As shown in the Appendix, uniqueness follows from the fact that if B and B' are K-reachable, then so is $B \cap B'$. Hence, $B_{K,\mathcal{C}}$ is the intersection of all K-reachable states and, in fact, the finest K-reachable belief state: $B_{K,\mathcal{C}} \subseteq B$ for all K-reachable states B.[6]

If the agent fails to deduce $g_{\mathcal{C}}$ (that is, if $B_{K,\mathcal{C}} \neq \{g_{\mathcal{C}}\}$), then, from his perspective, the contract is ambiguous: there are actions $a \in A$ such that $b_{K,\mathcal{C}}(a)$ contains two

[6] One may equivalently define $B_{K,\mathcal{C}}$ to be the (unique) K-reachable state B^* such that $B^* \subseteq B$ for all K-reachable states B. I am grateful to an anonymous referee for suggesting the present formulation.

or more outcomes. To close the model, an assumption regarding the agent's attitude toward such (perceived) ambiguity is required.

ASSUMPTION 1 (Ambiguity Aversion): *Given a contract $\mathcal{C}$, an agent of ability K and type θ evaluates actions $a \in A$ by the formula*

$$\begin{aligned} U_\theta(a,K,\mathcal{C}) &:= \min_{x \in b_{K,\mathcal{C}}(a)} u_\theta(x) \\ &= \min_{g \in B_{K,\mathcal{C}}} u_\theta\big(g(a)\big), \end{aligned}$$

and participates if and only if $\max_{a \in A} U_\theta(a,K,\mathcal{C}) \geq u_\theta(\bar{x}_\theta)$.

That is, the agent adopts a worst-case (maxmin) criterion when evaluating actions under beliefs $b_{K,\mathcal{C}}$. This is an extreme degree of ambiguity aversion, but many insights generated by the model hold under alternative assumptions: see Section IIIA.[7]

I conclude this section with an explicit example of a contract and an illustration of the cognitive procedure. Variations of this example will be used throughout the paper.

Example 1: A manager is recruiting an employee to work on a new project. The project involves several possible tasks (numbered 1 to 6) and employee types indicate their interest level in the project (Low or High). A (direct) mechanism consists of a pair of numbers (L,H) indicating the task number assigned based on the type report. Thus, $A = \Theta = \{L,H\}$ and $X = \{1, \ldots, 6\}$. Consider the contract consisting of the following five clauses:

C_1 : Exactly one type receives an even-numbered task.

C_2 : If H is even, then L is even.

C_3 : $L + H$ is either 3, 7, or 11.

C_4 : If $L \geq 5$ or $H \leq 2$, then $L > 3$ and $H < 4$.

C_5 : If $L \geq 5$ or $H \geq 4$, then $L > 3$ and $H > 2$.

An agent of ability $K \geq 2$ can combine C_1 and C_2 to deduce that H is odd and L is even. Hence, such agents can transition to state B where $b^B(L) = \{2,4,6\}$ and $b^B(H) = \{1,3,5\}$. Further refinement of these beliefs requires $K \geq 3$ because no pair of clauses eliminates any outcomes from this correspondence.[8] Only by combining C_3, C_4, *and* C_5 (simultaneously) with B can a new belief correspondence

[7] In this setup, ambiguity aversion can be interpreted as the attitude of an agent who is aware of his cognitive limitation and skeptical of the principal's motives: the fact that he cannot pin down the true mechanism raises suspicion that the principal is trying to deceive him. Only the worst-case criterion protects agents from bad outcomes (those dominated by their outside options). Thus, in the presence of potential manipulators, the worst-case criterion may be an advantageous heuristic for cognitively constrained individuals.

[8] For example, $\{(6,1),(4,3),(2,5)\} \subseteq C_3 \cap C_4$, so that no even number is eliminated for L and no odd number is eliminated for H even after combining C_3 and C_4 with beliefs B. A similar property holds for all other pairs of clauses.

be reached. In fact, performing this calculation reveals the true mechanism. Thus, ability $K \geq 3$ deduces $(L,H) = (4,3)$, $K = 2$ remains in state B above, and $K = 1$ learns nothing.

II. Implementation via Complex Contracts

A. *K-Implementation*

The principal seeks to implement a function $f : \Theta \to X$ specifying an outcome for each type. To do so, she designs a contract that *K-implements f*.

DEFINITION 4 (K-Implementation): *Let $K \geq 1$. A contract, $\mathcal{C}$, K-implements the function f if there is a profile $(a_\theta)_{\theta\in\Theta}$ of actions such that, for all $\theta \in \Theta$ and $a' \in A$,*

(*i*) $U_\theta(a_\theta, K, \mathcal{C}) \geq U_\theta(a', K, \mathcal{C})$ (*Incentive Compatibility*),

(*ii*) $U_\theta(a_\theta, K, \mathcal{C}) \geq u_\theta(\bar{x}_\theta)$ (*Individual Rationality*), *and*

(*iii*) $g_{\mathcal{C}}(a_\theta) = f(\theta)$.

A function f is K-implementable if there exists a contract that K-implements f. If f is K-implementable for some K, then f is implementable.

A contract $\mathcal{C}$ implements f by inducing beliefs that make each type θ wish to participate (Individual Rationality) and take a recommended action a_θ (Incentive Compatibility) such that $g_{\mathcal{C}}(a_\theta) = f(\theta)$. Thus, type θ is led to believe that a_θ is an optimal response even though he may prefer a different action if he could deduce the true mechanism $g_{\mathcal{C}}$. Since induced beliefs depend on K, a contract that K-implements f need not achieve implementation for other abilities K'. However, as we shall see, it will be without loss to consider a class of contracts that K-implement a given function for all K up to some bound.

As is typical in mechanism design, I focus on direct mechanisms. A contract is *direct* if $A = \Theta$. If $\mathcal{C}$ is direct and the profile $a_\theta := \theta$ satisfies all requirements of Definition 4, then $\mathcal{C}$ *directly K-implements f*. That is, to achieve direct implementation, a contract must induce beliefs making truthful reporting appear optimal for all types. Note that, for direct contracts, the IC and IR conditions depend only on the induced belief correspondence; that is, beliefs $b_{K,\mathcal{C}}$ make agents of all types prefer to participate and report truthfully. Arbitrary correspondences b (or their associated belief states) will be called *incentive-compatible* if they satisfy these conditions. The restriction to direct contracts is justified by the following result.

THEOREM 1 (A Revelation Principle): *If a function is implementable, then it is implementable by a direct contract.*

Like the standard revelation principle, Theorem 1 simplifies the search for implementable functions by restricting attention to direct contracts. Importantly, Theorem 1 is silent regarding the range of K for which implementation is achieved. As shown in

Section IIIC, this range increases as the action space expands. Thus, the only loss in restricting attention to direct contracts is to limit the range of K for which implementation can be achieved; the set of implementable functions is not affected.

If the agent is fully rational, then a function f is implementable if and only if it is *trivial*: for all $\theta, \theta' \in \Theta$, $f(\theta) \succsim_\theta f(\theta')$ and $f(\theta) \succsim_\theta \bar{x}_\theta$. The goal is to determine when and how nontrivial functions can be implemented for boundedly rational agents.

Example 2: This is similar to Example 1, but with three types and four tasks. Thus, $A = \Theta = \{L, M, H\}$ and $X = \{1, 2, 3, 4\}$. Preferences are given by the following table (ordering best to worst for each $\succsim_\theta$):

$\succsim_L$	4	3	2	1
$\succsim_M$	1	3	4	2
$\succsim_H$	3	1	2	4

Types L and M have outside option 3 and H has no outside option (equivalently, $\bar{x}_H = 4$, his least-preferred outcome). Suppose the agent is fully rational, so that for any contract he deduces f (condition (iii) of Definition 4). Let f_θ denote $f(\theta)$ and write $f = (f_L, f_M, f_H)$. Then $f^1 = (4, 1, 3)$ is trivial; $f^2 = (2, 2, 2)$ violates *IR*; and $f^3 = (4, 3, 1)$ and $f^4 = (4, 3, 2)$ satisfy *IR* but not *IC*. As we shall see, with boundedly rational agents, f^1 and f^3 are implementable but f^2 and f^4 are not.

The next definition provides a condition fully characterizing the set of implementable functions. For each $\theta \in \Theta$ and $x \in X$, let $L_\theta(x) := \{y \in X : x \succ_\theta y\}$ denote the strict lower contour of x under preferences $\succsim_\theta$.

DEFINITION 5 (IR-Dominance): *A function, f, is IR-Dominant if, for all $\theta, \theta' \in \Theta$ and all $x \in X$, $f(\theta) \succsim_\theta x$ if $L_{\theta'}(\bar{x}_{\theta'}) \supseteq L_\theta(x)$. Let $D(\bar{x})$ denote the set of all IR-Dominant functions.*

IR-Dominance requires type θ to prefer $f(\theta)$ over x whenever $\succsim_\theta$ and some other $\succsim_{\theta'}$ indicate that any outcome dominated by x (according to $\succsim_\theta$) is also dominated by $\bar{x}_{\theta'}$, the outside option for θ'. Intuitively, IR-Dominance expresses two necessary conditions for the existence of incentive-compatible beliefs. First, the worst-case outcome from truthful reporting must be at least as attractive as the outside option. Second, if satisfying the first condition makes type θ expect a worst-case outcome at least as good as x from misreporting as θ', then $f(\theta)$ must be even more attractive than x. In addition to characterizing the set of implementable functions, the concept of IR-Dominance is used to define the following class of contracts.

DEFINITION 6 (Complex Contract): *Let $f \in D(\bar{x})$. The (direct) contract $\mathcal{C}_f$ defined by*

$$\mathcal{C}_f := \left\{ D(\bar{x}) \backslash \{g\} : g \in D(\bar{x}) \text{ and } g \neq f \right\}$$

is the complex contract for f.

Each clause of $\mathcal{C}_f$ is formed by taking the set $D(\bar{x})$ and removing a single mechanism $g \in D(\bar{x})$; cycling through all choices of $g \neq f$ yields $\mathcal{C}_f$. Thus, each clause allows the agent to deduce that f is IR-Dominant, but provides only slightly more information by ruling out a single IR-Dominant function. This maximizes the number of clauses that must be combined in order to improve upon the knowledge that f is IR-Dominant.

As shown in Section IIB, the set $D(\bar{x})$ qualifies as an incentive-compatible belief state: beliefs $B = D(\bar{x})$ make every type θ wish to participate and respond truthfully. Thus, $\mathcal{C}_f$ can be interpreted as the result of a simple design heuristic: *minimize the informativeness of each clause* (*thereby maximizing the difficulty of performing transitions*) *subject to the constraint that each clause makes truthful reporting appear optimal.*

THEOREM 2: *If a function is implementable, then it is IR-Dominant. Moreover, there exists an integer $\bar{K}$ such that any nontrivial, IR-Dominant function f is directly K-implementable if and only if $K < \bar{K}$; in particular, $\mathcal{C}_f$ directly K-implements f for all $K < \bar{K}$.*

Theorem 2, the main result of this paper, establishes that IR-Dominance is necessary and (almost) sufficient for K-implementability: a nontrivial function is K-implementable if and only if it is IR-Dominant and $K < \bar{K}$. Moreover, the contract $\mathcal{C}_f$ achieves implementation for all $K < \bar{K}$, making it an "optimal" contract from the principal's perspective: if $\mathcal{C}_f$ does not K-implement f, neither does any other contract. Thus, a sophisticated principal has a strong incentive to introduce excess (but constrained) complexity into contracts.

The contract $\mathcal{C}_f$ exhibits a high degree of robustness not only to variation in K, but to variations on the cognitive procedure itself. This holds because (i) successful processing of *any* clause $C \in \mathcal{C}_f$ will transition the agent to the (incentive-compatible) state $D(\bar{x})$, and (ii) even from state $D(\bar{x})$, at least $\bar{K}$ clauses must be combined to reach a finer state. This makes state $D(\bar{x})$ easy to reach but difficult to escape, so that implementation is likely to be achieved even if the agent's cognitive process deviates from the specific model introduced here. For example, the agent need not perform multiple rounds of transitions or even understand all clauses of $\mathcal{C}_f$: randomly selecting any set of at most $\bar{K} - 1$ clauses for processing would result in beliefs $D(\bar{x})$. As long as the agent successfully processes at least one clause (but never $\bar{K}$ or more at once), implementation will be achieved. For additional extensions and robustness results, see Section III.

B. *An Illustration*

This section provides a sketch of the proof of Theorem 2 and also derives an explicit formula for the bound $\bar{K}$. For concreteness, the analysis is developed in the context of Example 2.

Two steps are needed to implement a nontrivial function f. The first is to derive an incentive-compatible belief state B, and the second is to construct a contract $\mathcal{C}$ inducing those beliefs. The contract must satisfy $g_{\mathcal{C}} = f$, forcing $f \in B$.

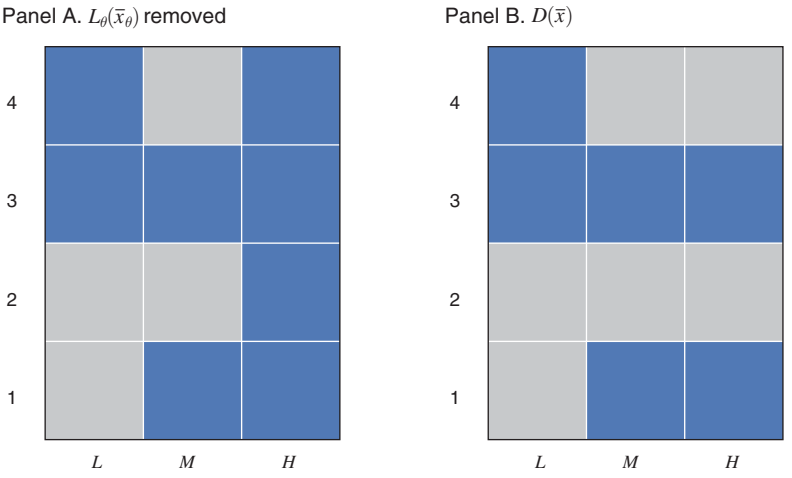

FIGURE 2. CONSTRUCTING $D(\bar{x})$

It turns out that every incentive-compatible B is a subset of $D(\bar{x})$, and that $D(\bar{x})$ itself is an incentive-compatible belief state. It follows that IR-Dominance is a necessary condition for implementability, because if $\mathcal{C}$ K-implements f by inducing beliefs $B_{K,\mathcal{C}} = B$, then $f = g_{\mathcal{C}} \in B \subseteq D(\bar{x})$.

The derivation of $D(\bar{x})$ is illustrated in Figure 2 using the framework and preferences of Example 2. The idea is to construct a maximal incentive-compatible correspondence, b^*, in two steps. First, in order for a correspondence to be incentive-compatible, it must satisfy the IR constraint. In panel A, outcomes strictly dominated by $\bar{x}_\theta$ (according to preferences $\succsim_\theta$) are eliminated as possible consequences of reporting θ. This makes the correspondence satisfy IR, but it violates IC: type H would prefer to misreport as type M. Therefore, in panel B, outcomes 2 and 4 are removed as possible consequences of report H, the minimal change needed to satisfy IC. By the maxmin criterion, IR is still satisfied. It is then a straightforward exercise to verify that $D(\bar{x})$ consists of all functions contained in b^*. In this case, $D(\bar{x})$ is represented by the correspondence b^* where $b^*(L) = \{3,4\}$ and $b^*(M) = b^*(L) = \{1,3\}$, as illustrated in panel B of Figure 2.

Now let $f \in D(\bar{x})$ and consider $\mathcal{C}_f$. Every clause of $\mathcal{C}_f$ is a subset of $D(\bar{x})$, and therefore the agent can reach state $D(\bar{x})$ by processing any clause. In order to reach a finer belief state, the agent must eliminate some outcome as a possible consequence of some action. This requires eliminating all functions in $D(\bar{x})$ passing through a particular point of the correspondence. For example, to eliminate 4 as a possible consequence of report L, the agent must eliminate the $2 \cdot 2 = 4$ functions in $D(\bar{x})$ passing through the point $(L,4)$. This requires $K \geq 4$ since each clause of $\mathcal{C}_f$ eliminates only a single function.

A simple induction argument establishes that if the agent is sophisticated enough to perform one such elimination, then he is sophisticated enough to deduce (through a series of K-valid transitions) the true mechanism $g_{\mathcal{C}} = f$. Thus, $\mathcal{C}_f$ has a "bang-bang" nature: the agent either deduces f or remains stuck in state $D(\bar{x})$.

How sophisticated must the agent be in order to perform an elimination? Observe that the level K needed to eliminate the point (θ,x) from the correspondence b^* is $\prod_{\theta'\neq\theta}|b^*(\theta')|$, because this is the number of functions in $D(\bar{x})$ passing through (θ,x). Thus, the *minimum* level K needed to refine b^* is

$$\bar{K} \coloneqq \min_{\theta\in\Theta}\prod_{\theta'\neq\theta}|b^*(\theta')|, \tag{2}$$

which is the sought-after formula for $\bar{K}$. An agent with $K<\bar{K}$ will end up in belief state $D(\bar{x})$, while $K\geq\bar{K}$ deduces the true mechanism $g_{\mathcal{C}}=f$. Thus, if f is nontrivial and $K\geq\bar{K}$, $\mathcal{C}_f$ fails to K-implement f. As shown in the Appendix, this implies no contract can K-implement f. Roughly, this follows from the fact that (i) any implementing contract must induce incentive-compatible beliefs $B\subseteq D(\bar{x})$, and (ii) $\mathcal{C}_f$ maximizes the level K need to refine beliefs $D(\bar{x})$. Thus, if some contract K-implements f, so must $\mathcal{C}_f$.

To conclude this section, the following example provides some additional insight into the structure of $\mathcal{C}_f$.

Example 3 (**Continued from Example 2**): Let $X=\{1,2,3,4\}$ and $A=\Theta=\{L,M,H\}$. Preferences and outside options are as in Example 2. Consider the contract, $\mathcal{C}$, consisting of the following eight clauses:

C_0 : M and H are odd, and $L\geq 3$.

C_1 : $L+M+H<10$.

C_2 : If $M+H=L$, then $H=1$.

C_3 : If $M=H=1$, then $L=3$.

C_4 : If $M=H=3$, then $L=4$.

C_5 : If $L=M$, then $H=3$.

C_6 : If $L=H$, then $M=3$.

C_7 : $L+M+H>5$.

Clause C_0 yields beliefs B where $b^B(L)=\{3,4\}$ and $b^B(M)=b^B(H)=\{1,3\}$, as in panel B of Figure 2. Thus, beliefs B coincide with $D(\bar{x})$ and are incentive compatible. Only ability $K\geq 4$ can refine these beliefs. In particular, combining clauses C_4–C_7 with B reveals $L=4$. From B, one could alternatively combine C_2, C_3, C_6, and C_7 to deduce $M=3$, or C_1, C_3, C_4, and C_6 to deduce $H=1$. No other combinations of four or fewer clauses allow any outcomes to be eliminated from the correspondence b^B. Thus, ability $K\geq 4$ deduces $f=(4,3,1)$ while abilities $1\leq K\leq 3$ remain stuck in state B. Since f is not trivial, implementation is achieved only for $K\leq 3$ (that is, $\bar{K}=4$).

This contract is equivalent to $\mathcal{C}_f$ in terms of induced beliefs ($b_{K,\mathcal{C}}=b_{K,\mathcal{C}_f}$ for all K) but is not actually $\mathcal{C}_f$. To construct $\mathcal{C}_f$, replace C_i ($i=1,\ldots,7$) with $C_0\cap C_i$. Essentially, this appends the statement "$L\in\{3,4\}$ and $M,H\in\{1,3\}$" to each C_i. Thus, in $\mathcal{C}_f$, every single clause allows the agent to transition to state $B=D(\bar{x})$, whereas in $\mathcal{C}$ the agent must process C_0 in order to reach B.

This example also illustrates another dimension along which $\mathcal{C}_f$ is robust: not only is $K \geq 4$ required to refine beliefs $D(\bar{x})$, but only three of the $\binom{7}{4} = 35$ sets of four clauses enable transitions to finer states. Thus, even an agent of ability $K = 4$ will get stuck in $D(\bar{x})$ if he is not patient enough to examine most combinations of four clauses from $\mathcal{C}_f$.

C. *Comparative Statics*

The analysis so far has fixed the profile $\bar{x}$ of outside options and (by Theorem 1) restricted attention to direct mechanisms ($A = \Theta$). In this section, I show how the set of implementable functions and the bound $\bar{K}$ vary with $\bar{x}$ and the choice of action space A.

Outside Options.—Both the set of implementable functions $D(\bar{x})$ and the bound $\bar{K}$ vary with the profile $\bar{x}$ of outside options. In this section only, I will write $\bar{K}(\bar{x})$ to emphasize this dependency. The following result is a straightforward consequence of (the proof of) Theorem 2.

PROPOSITION 1: *If* $\bar{x}'_\theta \succsim_\theta \bar{x}_\theta$ *for all* θ, *then* $D(\bar{x}') \subseteq D(\bar{x})$ *and* $\bar{K}(\bar{x}') \leq \bar{K}(\bar{x})$.

In other words, the set of implementable functions shrinks and $\bar{K}$ decreases as outside options become more attractive for all types. Intuitively, this follows from formula (2) for $\bar{K}(\bar{x})$ and the fact that better outside options shrink the correspondence b^* associated with $D(\bar{x})$ (with better outside options, more outcomes must be eliminated in order to satisfy IR).

An interesting special case is when each $\bar{x}_\theta$ is the worst-possible outcome in X for type θ; that is, it is as if types do not have outside options at all. Then $D(\bar{x}) = G$, so that every function is implementable and $\bar{K}(\bar{x}) = |X|^{|\Theta|-1}$. This case still requires the full proof to establish both implementability as well as the optimality of complex contracts.[9]

Larger Action Sets.—By Theorem 1, the choice of action space A does not affect the set of implementable functions. However, the range of K for which a function can be implemented depends on A. In this section, I show that $\bar{K}$ increases as $|A|$ increases. Throughout, I assume $|A| \geq |\Theta|$.

Let $f \in D(\bar{x})$. The definition of $\mathcal{C}_f$ can be adapted to the action space A as follows. First, relabel elements to express A as a (disjoint) union $A = \Theta \cup A'$. Let $b^*: \Theta \rightrightarrows X$ denote the correspondence associated with $D(\bar{x})$. Extend this to a correspondence from A to X by letting $b^*(a) = X$ for all $a \in A \backslash \Theta$. Now choose any extension f^A of f to the domain A. Let $D^A(\bar{x}) = \{g \in G : g|_\Theta \in D(\bar{x})\}$ be the set of functions $g: A \to X$ that restrict to functions in $D(\bar{x})$ on the domain Θ, and consider the contract

$$\mathcal{C}_f^A := \{D^A(\bar{x}) \backslash \{g\} : g \in D^A(\bar{x}) \text{ and } g \neq f^A\}.$$

[9] Note that this case involves induced beliefs making the agent believe (via the maxmin criterion) that he will receive the worst-possible outcome of X by participating in the mechanism. If X contains extreme outcomes (e.g., large fines), then the agent likely has a more attractive outside option.

This is analogous to the contract $\mathcal{C}_f$, where clauses are of the form $D(\bar{x})\backslash\{g\}$ for $g \neq f$. By Theorems 1 and 2, IR-Dominance is a necessary condition for implementability (even with arbitrary action sets A). The next result, like Theorem 2, establishes a partial converse.

PROPOSITION 2: *Suppose* $|A| \geq |\Theta|$ *and that* $f \in D(\bar{x})$ *is nontrivial. The contract* $\mathcal{C}_f^A$ *K-implements f for all* $K < \bar{K}^A$, *where*

$$\bar{K}^A := \min_{a \in A} \prod_{a' \neq a} |b^*(a')|.$$

The logic of Proposition 2 is similar to that of Theorem 1. If $K < \bar{K}^A$, the agent gets stuck in belief state $D^A(\bar{x})$; if $K \geq \bar{K}^A$, he deduces the true mechanism. Beliefs $D^A(\bar{x})$ are incentive-compatible because (under the relabeling) actions $a \in A'$ are completely ambiguous and, hence, weakly dominated by actions $a \in \Theta$ (where outcomes coincide with those of b^*). Thus, implementation is achieved only for $K < \bar{K}^A$ if f is nontrivial.

Note that $\bar{K}^A$ is strictly increasing in the cardinality of A, so that (in the limit) only IR-Dominance matters. This suggests the principal may wish to inflate A indefinitely, thereby achieving implementation for any K she desires. In practice, the principal may be constrained by language needed to describe mechanisms or clauses on larger action spaces.

III. Extensions, Variations, and Robustness

A. *Ambiguity Attitude*

In this section, I show how to examine the model under alternative assumptions regarding the agent's attitude toward ambiguity or, more generally, under alternative assumptions about how the agent evaluates belief correspondences.

The procedure for forming beliefs $b_{K,\mathcal{C}}$ is independent of how the agent ranks actions (type reports) under those beliefs. Therefore, solving the model under alternative ambiguity assumptions requires two steps:

(i) Given a function f, find a belief correspondence, b, such that $f \in B^b$ and b satisfies appropriate IR and IC conditions under the alternative ambiguity assumption. If no such b exists, f cannot be implemented.

(ii) The contract

$$\mathcal{C}_f^b = \left\{B^b\backslash\{g\} : g \in B^b \text{ and } g \neq f\right\}$$

implements f for all $K < \bar{K}^b := \min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b(\theta')|$.

This procedure accommodates many different modeling assumptions, including some that do not necessarily regard the multi-valuedness of $b_{K,\mathcal{C}}$ as stemming from ambiguity. For example, one could assume the agent holds a prior on G which is updated (via Bayes' rule) to a posterior on $B_{K,\mathcal{C}}$ after processing $\mathcal{C}$. This way, the

agent assigns an expected utility to each action, and appropriate IR and IC constraints can be defined.

To maximize the range of K for which implementation can be achieved, choose a correspondence b from step (i) that maximizes $\bar{K}^b$. For maxmin agents, this is done by taking $b = b^*$ where $B^{b^*} = D(\bar{x})$. Different assumptions generally require different choices of b and, unlike the maxmin case, this choice may also depend on f. Thus, ambiguity attitude determines the set of implementable functions, but the contract $\mathcal{C}_f^b$ satisfies robustness and comparative static properties similar to those of $\mathcal{C}_f$.

Although $\mathcal{C}_f$ is derived under the assumption of maxmin preferences, it turns out to be robust to a wide range of ambiguity attitudes; in particular, those parametrized by the Hurwicz (1951) α-criterion. Under this criterion, an agent with parameter $\alpha \in [0,1]$, utility function u_θ, and beliefs b assigns utility $U_\theta^\alpha(a)$ to action a, where

$$U_\theta^\alpha(a) := \alpha \min_{x \in b(a)} u_\theta(x) + (1-\alpha) \max_{x \in b(a)} u_\theta(x). \tag{3}$$

At $\alpha = 1$, the agent is maxmin; at $\alpha = 0$, he is "maxmax" (he behaves as if the best possible outcome in $b(a)$ will attain). In general, cardinal properties of u_θ affect the agent's behavior under this criterion. Nonetheless, $\mathcal{C}_f$ still implements $f \in D(\bar{x})$ for all $K < \bar{K}$ and all $\alpha \in [0,1]$.

PROPOSITION 3: *If f is IR-Dominant and the agent has α-maxmin preferences, then $\mathcal{C}_f$ implements f for all $K < \bar{K}$.*

Note that, for a given value of α, the set of implementable functions (or the range of admissible K) may expand. Rather than a complete characterization (which would depend on α as well as cardinal properties of the functions u_θ), Proposition 3 should be understood as a robustness result: a principal who is uncertain about the agent's ambiguity attitude (but believes preferences are α-maxmin for some unknown α) can implement an IR-Dominant function for all $K < \bar{K}$ by choosing $\mathcal{C}_f$.

B. *Endogenous K*

In the baseline version of the model, the parameter K is a fixed attribute of the agent. In this section, I consider the possibility that K may respond to incentives. In particular, the agent may wish to acquire a higher ability K if doing so results in sharper beliefs and, hence, greater ability to manipulate the mechanism.

To attain ability K, the agent suffers a cost $c(K)$, where c is strictly increasing and satisfies $c(1) = 0$. One may interpret $c(K)$ as a cost (psychological or otherwise) of computational effort. Given c, the principal seeks to design a contract that *c-implements* her objective.

DEFINITION 7: *Let c be a cost function. A contract, $\mathcal{C}$, (directly) c-implements a function f if there exists K^* such that, for all θ*

$$K^* \in \operatorname*{argmax}_{K \geq 1} \max_{\theta' \in \Theta} U_\theta(\theta', K, \mathcal{C}) - c(K),$$

and for all $\theta, \theta' \in \Theta$,

(*i*) $U_\theta(\theta, K^*, \mathcal{C}) - c(K^*) \geq u_\theta(\bar{x}_\theta)$,

(*ii*) $U_\theta(\theta, K^*, \mathcal{C}) \geq U_\theta(\theta', K^*, \mathcal{C})$, *and*

(*iii*) $g_{\mathcal{C}}(\theta) = f(\theta)$.

A function f is c-implementable if there exists a contract that c-implements f.

Under c-implementation, an agent of type θ weighs the (anticipated) benefit $\max_{\theta' \in \Theta} U_\theta(\theta', K, \mathcal{C})$ of participating in the mechanism under beliefs $B_{K,\mathcal{C}}$ against the cost $c(K)$ of acquiring ability K.[10] For implementation to be achieved, an agent-optimal ability K^* must induce beliefs that satisfy the IR and IC constraints. For each θ and f, let $u^*_\theta(f) := \max_{\theta' \in \Theta} u_\theta(f(\theta'))$.

PROPOSITION 4: *If f is c-implementable, then f is IR-Dominant. If f is nontrivial and IR-Dominant, then $\mathcal{C}_f$ c-implements f if* $\max_{\theta \in \Theta} u^*_\theta(f) - \min_{g \in D(\bar{x})} u_\theta(g(\theta)) < c(\bar{K})$.

Proposition 4 establishes that IR-Dominance remains necessary and (almost) sufficient for implementability. The condition for $\mathcal{C}_f$ to be effective says that, for each type θ, the net payoff from acquiring ability $\bar{K}$ and perfectly manipulating the mechanism $(u^*_\theta(f) - c(\bar{K}))$ does not exceed the payoff from choosing $K = 1$ and ending up with (anticipated) payoff $U_\theta(\theta, K = 1, \mathcal{C}_f) = \min_{g \in D(\bar{x})} u_\theta(g(\theta))$. Intuitively, this is the relevant comparison because $\mathcal{C}_f$ induces beliefs $D(\bar{x})$ for all $K < \bar{K}$. Since c is strictly increasing, this means only $K = 1$ or $K = \bar{K}$ can be optimal for the agent. Thus, the condition ensures that $\mathcal{C}_f$ achieves implementation as long as each type of agent prefers $K = 1$ over $\bar{K}$, given c.

C. *Finer Belief States*

So far, the analysis has allowed variation in K but fixed the family $\mathcal{B}$ of belief states. In particular, a set $B \subseteq G$ belongs to $\mathcal{B}$ if and only if there is a correspondence b such that $B = B^b$. In this section, I consider more general families of beliefs, defined as follows.

DEFINITION 8 (Belief System): *A belief system is a family $\hat{\mathcal{B}}$ of subsets of G such that*

B1. every $B \in \hat{\mathcal{B}}$ is nonempty;

B2. if $B, B' \in \hat{\mathcal{B}}$ and $B \cap B' \neq \varnothing$, then $B \cap B' \in \hat{\mathcal{B}}$;

B3. if $B \in \mathcal{B}$, then $B \in \hat{\mathcal{B}}$.

[10] This assumes the agent correctly assesses the benefit from choosing K before acquiring that ability. The "circularity" of this approach has obvious conceptual drawbacks. Nonetheless, correct forecasting of this sort seems to be a natural benchmark. For more on costs and benefits of reasoning, see Alaoui and Penta (2016a, b).

Property B1 states that belief states are nonempty sets of mechanisms. Property B2 states that if the agent can recall that a mechanism satisfies one property, and also recall that it satisfies a second, then he can recall that it satisfies both properties. Finally, B3 states that the agent is able to recall the set of possible outcomes associated with each action. It is easy to see that $\mathcal{B}$ satisfies all three properties.

Some additional terminology is needed to define the cognitive procedure for arbitrary belief systems. A *cognitive type* is a pair $T = \left(K, \hat{\mathcal{B}}\right)$ where $K \geq 1$ is an integer and $\hat{\mathcal{B}}$ is a belief system. If $T' = \left(K', \hat{\mathcal{B}}'\right)$, then $T' \leq T$ means $K' \leq K$ and $\hat{\mathcal{B}}' \subseteq \hat{\mathcal{B}}$. Thus, type T is more sophisticated in that he has both greater computational ability and a richer set of belief states than type T'.

For any type $T = \left(K, \hat{\mathcal{B}}\right)$, the definitions of K-validity and K-reachability can be adapted from Definitions 1 and 2 by replacing $\mathcal{B}$ with $\hat{\mathcal{B}}$. Call the resulting concepts *T-validity* and *T-reachability*, respectively. Given these definitions, the concept of induced belief states (Definition 3) can be adapted as well. Properties B1 and B2 ensure Lemma 1 holds for arbitrary T (see the Appendix). Thus, given a contract $\mathcal{C}$, there is a unique induced belief state, denoted $B_{T,\mathcal{C}}$. The *effective* belief state, denoted $B^*_{T,\mathcal{C}}$, is the smallest member of $\mathcal{B}$ containing $B_{T,\mathcal{C}}$. Hence, the effective belief state is associated with a correspondence $b^*_{T,\mathcal{C}}$ given by

$$b^*_{T,\mathcal{C}}(a) \coloneqq \left\{g(a) : g \in B^*_{T,\mathcal{C}}\right\} = \left\{g(a) : g \in B_{T,\mathcal{C}}\right\}.$$

The idea of an effective belief state is that if the agent has arrived in state $B_{T,\mathcal{C}}$ and if $g \in B_{T,\mathcal{C}}$, then he considers $g(a)$ to be a possible consequence of action a. Thus, it is as if his beliefs are represented by $b^*_{T,\mathcal{C}}$ and, hence, the state $B^*_{T,\mathcal{C}} \in \mathcal{B}$.

Once again, an agent of cognitive type T evaluates his belief by the maxmin criterion. This is equivalent to evaluating his effective belief by the maxmin criterion. Hence, the definition of K-implementability can be extended to T-implementability in the obvious way. For ease of exposition, I restrict attention to direct contracts $(A = \Theta)$.

Given a contract $\mathcal{C}$, a cognitive type T, and a function $f \in B^*_{T,\mathcal{C}}$, let

$$\mathcal{C}_{T,f} \coloneqq \left\{B^*_{T,\mathcal{C}} \backslash \{g\} : g \in B^*_{T,\mathcal{C}}, g \neq f\right\}.$$

This is similar to the complex contract $\mathcal{C}_f$, but replaces $D(\bar{x})$ with $B^*_{T,\mathcal{C}}$. Each clause indicates that $f \in B^*_{T,\mathcal{C}}$, but eliminates only one function from $B^*_{T,\mathcal{C}}$.

PROPOSITION 5: *If a contract, $\mathcal{C}$, T-implements a function f, then f is IR-Dominant and $\mathcal{C}_{T,f}$ T'-implements f for all $T' \leq T$.*

The logic of Proposition 5 is similar to that of Theorem 2. For a function to be implementable, it must be contained in an incentive-compatible correspondence and, hence, IR-Dominant. The main difference is that some choices of $\hat{\mathcal{B}}$ may make the agent highly adept at transitioning away from state $D(\bar{x})$, and therefore $\mathcal{C}_f$ may fail to implement some IR-Dominant f.[11] But if a contract, $\mathcal{C}$, T-implements f by

[11] In particular, under $\mathcal{C}_f$, the agent may arrive at beliefs $B \subsetneq D(\bar{x})$ that are not incentive compatible.

inducing (effective) beliefs $B^*_{T,\mathcal{C}}$, then $\mathcal{C}_{T,f}$ T'-implements f for all $T' \leq T$ because $\mathcal{C}_{T,f}$ maximizes the difficulty of escaping the (incentive-compatible) state $B^*_{T,\mathcal{C}}$.

IV. Discussion

A. *Related Literature*

A growing literature on behavioral mechanism design has emerged with the goal of understanding how various departures from standard rational behavior influence the design and effectiveness of mechanisms. In one branch, agents understand game forms and mechanisms but exhibit nonstandard choice or strategic behavior. For example, Korpela (2012) and de Clippel (2014) study implementation for agents with nonstandard choice functions, while de Clippel, Saran, and Serrano (2018) and Kneeland (2018) study mechanism design for agents with level-k strategic reasoning (Stahl and Wilson 1994, 1995).[12] The literature on mechanism design with ambiguity-averse agents (Gilboa and Schmeidler 1989) also belongs to this category. Bose and Renou (2014) argues that a designer cannot benefit from introducing ambiguity into the allocation rule unless a correspondence (rather than a function) is to be implemented, and construct a mechanism inducing endogenous ambiguity about the types of other players. In contrast, my results show that *perceived* ambiguity about the allocation rule can help the designer achieve her goals: the principal specifies a complete, unambiguous mechanism, but agents misperceive the rule to be ambiguous, to the principal's advantage.[13]

Another, less-developed branch considers the possibility that agents, independently of their strategic reasoning ability or other psychological traits, may not fully understand mechanisms presented to them. That is, they may hold incorrect or incomplete beliefs about how action profiles map to outcomes. The main challenge of this avenue is that it requires new models of bounded rationality indicating how agents form beliefs or make decisions when confronted with complex mechanisms. This paper develops such a model based on the idea that the ability to combine different pieces of information (and retain new facts derived in the process) is a key determinant of problem-solving ability. Consequently, the agent is sensitive to the way information is framed.[14]

As part of the second branch, this paper is most closely related to a pair of papers by Glazer and Rubinstein (2012, 2014)—henceforth, GR12/14. These papers study persuasion with boundedly rational agents: all agents (regardless of type) wish to be accepted by the principal, but the principal only wants to accept a particular subset of types. The papers differ in the manner in which agents are bounded as well as the implementation objective faced by the principal. In GR12, the principal specifies a set of conditions (each required to take a particular syntactical form) necessary for

[12] See also Koszegi (2014) for a recent survey of the behavioral contracting literature.

[13] Di Tillio, Kos, and Messner (2016) shows that a seller can benefit from using an ambiguous mechanism when buyers are ambiguity averse. For more on mechanism design with ambiguity aversion, see Bodoh-Creed (2012); Bose, Ozdenoren, and Pape (2006); Bose and Daripa (2009); and Wolitzky (2016).

[14] Salant and Rubinstein (2008) studies a general model where the framing of alternatives (not information) influences choice behavior, and Salant and Siegel (2018) applies this framework to a contracting model where a seller seeks to influence buyers through framing.

acceptance. The agent, rather than forming beliefs and acting on them, adheres to a particular algorithm for constructing a response. Crucially, the procedure initializes at the true type and is defined using the syntactical structure of the conditions. In GR14, the principal asks the agent a series of questions about his type and agents have limited ability to detect patterns in the set of acceptable responses. The same syntactical structure is needed to define the patterns that agents detect. The principal solves a constrained implementation problem where all truthful, acceptable types must be accepted while minimizing the probability that manipulators are accepted (manipulators can lie about their type; truth-tellers cannot). They show that this probability depends only on the number of acceptable types and that it decreases very quickly as the set of acceptable types expands.

Like GR12/14, this paper introduces a novel concept of bounded rationality and applies it in a principal-agent setting. However, the model and results differ in several ways. First, I study an implementation problem involving an arbitrary number of outcomes, heterogeneous preferences, and outside options. The principal's implementation objective is standard and is not subject to any particular constraints on form or content.[15] Second, agents in my model are bounded in a different way: they are limited in their ability to combine different pieces of information, and for this reason I abstract away from syntactical details of the contracting environment. Finally, the implementation results presented here are qualitatively different from those of GR12 and GR14. Implementation is deterministic, and the main results show that well-crafted complex contracts are robust to a variety of cognitive types and procedures.

The issue of robustness to nonstandard agent behavior has received some attention in the literature. Eliaz (2002), for example, considers an implementation setting where some players are error-prone and the designer seeks a mechanism robust to this possibility, while Li (2017) proposes an implementation concept robust to imperfect strategic reasoning in extensive-form games. A key result of this paper is that when cognitive ability (affecting the agent's perception of the game form) is the dimension of interest, strong robustness results emerge "for free": any goal that can be achieved through exploitation of limited cognitive ability can be achieved in a way that is highly robust to heterogeneity in cognitive abilities and procedures.

B. *Conclusion*

This paper has studied a mechanism design problem involving a principal and a single, boundedly rational agent. By designing contracts to exploit the agent's limited cognitive ability, the principal can implement a large class of objective functions (those satisfying a simple IR-Dominance condition) provided the agent is not too sophisticated. Without loss of generality, the principal adheres to a simple design principle: minimize the informativeness of each clause subject to the constraint that each clause makes truthful reporting appear optimal. Consequently, the optimal contract is highly robust to heterogeneity in cognitive ability as well as

[15] In particular, no syntactical structure is imposed and, like GR12/14, there are no costs associated with designing longer contracts. Introducing such costs, as in Battigalli and Maggi (2002), may be an interesting avenue for future research.

several variations on the agent's cognitive procedure. The analysis is grounded in a novel framework for bounded rationality where imperfect memory and computational ability limit the agent's ability to solve problems.

The model of cognition introduced in this paper is neither formally nor conceptually bound to the domain of implementation theory. It can be reformulated, for example, as a general model of non-Bayesian information processing. Let Ω denote a set of states and $\hat{\mathcal{B}}$ a family of nonempty subsets of Ω closed under nonempty intersections. Suppose an agent is presented with a set $\mathcal{F} = \{E_1, \ldots, E_n\}$ of events $E_i \subseteq \Omega$ such that $\bigcap_{E_i \in \mathcal{F}} E_i \neq \varnothing$. For example, each E_i could represent the realization of a signal (from a partitional information structure) indicating that the true state belongs to E_i. Alternatively, $\mathcal{F}$ could be interpreted as a *frame* for the event $E = \bigcap_{E_i \in \mathcal{F}} E_i$ (that is, E is framed as a set of events that jointly pin down E, similar to the way a contract is a set of clauses pinning down a mechanism). The family $\hat{\mathcal{B}}$ represents a set of feasible belief states for the agent. For any $K \geq 1$, the cognitive procedure for processing $\mathcal{F}$ can be adapted from the general model presented in Section IIIC, providing an intuitive and portable theory of complexity in information processing. Further development of this framework and its applications may be an interesting avenue for future research.

Appendix A: Proofs

A. *Preliminaries*

This section establishes some basic properties of the cognitive procedure. Since Proposition 5 utilizes the more general model introduced in Section IIIC, results are presented for general cognitive types $T = \left(K, \hat{\mathcal{B}}\right)$ where $\hat{\mathcal{B}}$ is a belief system satisfying properties B1–B3 of Definition 8. Throughout, $\mathcal{B}$ denotes the baseline belief system where $B \in \mathcal{B}$ if and only if there is a correspondence b such that $B = B^b$.

Given $T = \left(K, \hat{\mathcal{B}}\right)$, a transition $B \overset{\mathcal{C}'}{\rightarrow} B'$ is *T-valid* (under contract $\mathcal{C}$) if $B, B' \in \hat{\mathcal{B}}$, $\mathcal{C}' \subseteq \mathcal{C}$ with $|\mathcal{C}'| \leq K$, and

$$B \cap \left(\bigcap_{C \in \mathcal{C}'} C \right) \subseteq B'.$$

A state $B \in \hat{\mathcal{B}}$ is *T-reachable* if there is a sequence

$$G = B^0 \overset{\mathcal{C}^1}{\rightarrow} B^1 \overset{\mathcal{C}^2}{\rightarrow} B^2 \overset{\mathcal{C}^3}{\rightarrow} \cdots \overset{\mathcal{C}^n}{\rightarrow} B^n = B$$

of T-valid transitions. It is easy to see that K-validity and K-reachability (Definitions 1 and 2) are special cases of T-validity and T-reachability, respectively, where $T = (K, \mathcal{B})$.

LEMMA A.1: *Let $\mathcal{C}$ be a contract, $T = \left(K, \hat{\mathcal{B}}\right)$, and $B, B' \in \hat{\mathcal{B}}$. Then,*

(*i*) *if $B \overset{\mathcal{C}'}{\rightarrow} B'$ is T-valid and $B' \subseteq B'' \in \hat{\mathcal{B}}$, then $B \overset{\mathcal{C}'}{\rightarrow} B''$ is T-valid;*

(*ii*) *if B and B' are T-reachable, then $B \cap B' \neq \varnothing$. Hence, $B \cap B' \in \hat{\mathcal{B}}$.*

PROOF:

For (i), observe that if $B \xrightarrow{\mathcal{C}'} B'$ is T-valid, then $B \cap \left(\bigcap_{C\in\mathcal{C}'} C\right) \subseteq B' \subseteq B''$. Thus, $B \xrightarrow{\mathcal{C}'} B''$ is T-valid. For (ii), observe that $\bigcap_{C\in\mathcal{C}} = \{g_{\mathcal{C}}\}$. Therefore, $g_{\mathcal{C}} \in \bigcap_{C\in\mathcal{C}'} C$ for all nonempty $\mathcal{C}' \subseteq \mathcal{C}$. Now suppose B is T-reachable. Then there is a sequence

$$G = B^0 \xrightarrow{\mathcal{C}^1} B^1 \xrightarrow{\mathcal{C}^2} B^2 \xrightarrow{\mathcal{C}^3} \cdots \xrightarrow{\mathcal{C}^n} B^n = B$$

of T-valid transitions. If $g_{\mathcal{C}} \in B^i$ for some i, then $g_{\mathcal{C}} \in B^i \cap \left(\bigcap_{C\in\mathcal{C}^i} C\right) \subseteq B^{i+1}$, so that $g_{\mathcal{C}} \in B^{i+1}$. Since $g_{\mathcal{C}} \in G = B^0$, it follows that $g_{\mathcal{C}} \in B$. Thus, if B and B' are T-reachable, $B \cap B' \neq \varnothing$. By B2, this implies $B \cap B' \in \hat{\mathcal{B}}$. ∎

DEFINITION A.1: *Let $\mathcal{C}$ be a contract, $T = \left(K, \hat{\mathcal{B}}\right)$, and $B, B' \in \hat{\mathcal{B}}$ such that $B' \subseteq B$. Then B' is T-reachable from B if there exists a sequence*

$$B = B^0 \xrightarrow{\mathcal{C}^1} B^1 \xrightarrow{\mathcal{C}^2} B^2 \xrightarrow{\mathcal{C}^3} \cdots \xrightarrow{\mathcal{C}^n} B^n = B'$$

of T-valid transitions where $B^i \subseteq B$ for all i.

Notice that T-reachability is a special case of Definition A.1 (take $B = G$ for T-reachability). Also, if B' is T-reachable from B and $B' \xrightarrow{\mathcal{C}'} B''$ is T-valid, then B'' is T-reachable from B. Thus, if B is T-reachable and B' is T-reachable from B, then B' is T-reachable.

LEMMA A.2: *If $B, B' \in \hat{\mathcal{B}}$ are T-reachable, then $B \cap B'$ is T-reachable from B.*

PROOF:

Since B' is T-reachable, there is a sequence

$$G = \hat{B}^0 \xrightarrow{\mathcal{C}^1} \hat{B}^1 \xrightarrow{\mathcal{C}^2} \hat{B}^2 \xrightarrow{\mathcal{C}^3} \cdots \xrightarrow{\mathcal{C}^n} \hat{B}^n = B'$$

of T-valid transitions. Observe that

$$B \cap \left(\bigcap_{C\in\mathcal{C}^1} C\right) \subseteq \hat{B}^0 \cap \left(\bigcap_{C\in\mathcal{C}^1} C\right) \subseteq \hat{B}^1 \quad \text{and} \quad B \cap \left(\bigcap_{C\in\mathcal{C}^1} C\right) \subseteq B.$$

Thus,

$$B \cap \left(\bigcap_{C\in\mathcal{C}^1} C\right) \subseteq B \cap \hat{B}^1.$$

It follows that $B \xrightarrow{\mathcal{C}'} B \cap \hat{B}^1$ is a T-valid transition (note that $B \cap \hat{B}^1 \in \hat{\mathcal{B}}$ by Lemma A.1). If $n = 1$, then $\hat{B}^1 = B'$ and there is nothing left to prove. So, suppose $n > 1$. Proceeding by induction, suppose $1 \leq i < n$ and that $B \xrightarrow{\mathcal{C}^1} B \cap \hat{B}^1 \xrightarrow{\mathcal{C}^2} B \cap \hat{B}^1 \cap \hat{B}^2 \xrightarrow{\mathcal{C}^3} \cdots \xrightarrow{\mathcal{C}^i} B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^i$ is a sequence of T-valid transitions. Then

$$B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^i \cap \left(\bigcap_{C\in\mathcal{C}^{i+1}} C\right) \subseteq \hat{B}^i \cap \left(\bigcap_{C\in\mathcal{C}^{i+1}} C\right) \subseteq \hat{B}^{i+1}.$$

Thus,

$$B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^i \cap \left(\bigcap_{C \in \mathcal{C}^{i+1}} C\right) \subseteq B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^i \cap \hat{B}^{i+1}$$

and $B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^i \xrightarrow{\mathcal{C}^{i+1}} B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^{i+1}$ is a T-valid transition. By induction, then, $B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^n$ is T-reachable from B (let $B^i := B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^i$ for all $i = 1, \ldots, n$; clearly, $B^i \subseteq B$ for all i). Since $\hat{B}^n = B'$, it follows that $B \cap \hat{B}^1 \cap \cdots \cap \hat{B}^n \subseteq B \cap B'$, and so $B \cap B'$ is T-reachable from B by Lemma A.1. ∎

LEMMA A.3: *For every contract $\mathcal{C}$ and type T, there exists a unique induced belief state $B_{T,\mathcal{C}}$.*

PROOF:

Fix $\mathcal{C}$ and T and suppose B and $\hat{B}$ are induced belief states. By Lemma A.2, $B \cap \hat{B}$ is T-reachable from B. Thus, there is a sequence of T-valid transitions $B = B^0 \xrightarrow{\mathcal{C}^1} B^1 \xrightarrow{\mathcal{C}^2} B^2 \xrightarrow{\mathcal{C}^3} \cdots \xrightarrow{\mathcal{C}^n} B^n = B \cap \hat{B}$ such that $B^i \subseteq B$ for all i. If $B \cap \hat{B} \subsetneq B$, then there is a smallest $i^* \geq 1$ such that $B^{i^*} \subsetneq B$. But then the transition $B = B^{i^*-1} \xrightarrow{\mathcal{C}^{i^*}} B^{i^*}$ is T-valid, contradicting the fact that B is an induced belief state. Thus, $B = B \cap \hat{B}$. A similar argument establishes that $\hat{B} = B \cap \hat{B}$. Thus, $B = \hat{B}$, as desired. ∎

LEMMA A.4: *Let $\mathcal{C}$ be a contract and $T = \left(K, \hat{\mathcal{B}}\right)$. The state $B_{T,\mathcal{C}}$ is the (unique) T-reachable state $B^* \in \hat{\mathcal{B}}$ such that $B^* \subseteq B$ for all T-reachable states B.*

PROOF:

By Lemma A.2, $B \cap B'$ is T-reachable whenever B and B' are T-reachable. Let B^* be the intersection of all T-reachable states (this is a finite intersection because $\hat{\mathcal{B}}$ is finite). By Lemma A.2, B^* is T-reachable. By construction, $B^* \subseteq B$ for all T-reachable states B. Thus, if some other T-reachable state B' satisfies $B' \subseteq B$ for all T-reachable states B, we have $B^* \subseteq B'$ and, therefore, $B' = B^*$. To see that $B_{T,\mathcal{C}} = B^*$, it will suffice (by Lemma A.3) to show that B^* is an induced belief state. Suppose $B^* \xrightarrow{\mathcal{C}'} B'$ is T-valid. Then B' is T-reachable, forcing $B^* \subseteq B'$ by definition of B^*. Thus, B^* is an induced belief state. ∎

LEMMA A.5: *If B is T-reachable, then $B_{T,\mathcal{C}}$ is the intersection of all states $\hat{B} \in \hat{\mathcal{B}}$ that are T-reachable from B. In particular, $B_{T,\mathcal{C}}$ is the intersection of all T-reachable states.*

PROOF:

Let B be a T-reachable state. As shown in the proof of Lemma A.4, $B_{T,\mathcal{C}}$ is the intersection of all T-reachable states. Clearly, this coincides with the intersection of all sets of the form $B \cap B'$ where B' is T-reachable. To complete the proof, we show that a state $\hat{B}$ is T-reachable from B if and only if it is of the form $\hat{B} = B \cap B'$ for some T-reachable B'. If B' is T-reachable, then (by Lemma A.2), the set $B \cap B'$ is T-reachable from B. Conversely, suppose $\hat{B}$ is T-reachable from B. Then, by

definition, $\hat{B} \subseteq B$. Moreover, $\hat{B}$ is T-reachable because B is T-reachable. Take $B' = \hat{B}$; then $\hat{B} = B \cap B'$. ∎

B. *Proof of Theorems 1 and 2*

For any $Y \subseteq X$ and $\theta \in \Theta$, let $L_\theta(Y)$ denote the largest (possibly empty) strict lower-contour set of $\succsim_\theta$ contained in Y. That is, there exists $x \in X$ such that $L_\theta(Y) = L_\theta(x)$ and, for all $x' \in X$, $L_\theta(x') \subseteq Y \Rightarrow L_\theta(x') \subseteq L_\theta(Y)$. Clearly, any two sets $L_\theta(Y), L_\theta(Y')$ are ordered by set inclusion.

Let L_θ^* be the largest set of the form $L_\theta(Y)$ subject to $Y = L_{\theta'}(\bar{x}_{\theta'})$ $(\theta' \in \Theta)$. That is, there exists θ' such that $L_\theta^* = L_\theta\big(L_{\theta'}(\bar{x}_{\theta'})\big)$ and, for all θ'', $L_\theta\big(L_{\theta''}(\bar{x}_{\theta''})\big) \subseteq L_\theta^*$. The set L_θ^* is well defined because there are only finitely many sets of the form $Y = L_{\theta'}(\bar{x}_{\theta'})$, and contour sets $L_\theta(Y)$ are linearly ordered by set inclusion.

Define $b^*: \Theta \rightrightarrows X$ by $b^*(\theta) := X \backslash L_\theta^*$ and let $\bar{K} := \min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b^*(\theta')|$. Note that $b^*(\theta) \neq \varnothing$ for all θ because no strict lower contour set contains all of X. Thus, $\bar{K} \geq 1$.

LEMMA A.6: *The set of IR-Dominant functions, $D(\bar{x})$, satisfies the following:*

(*i*) $D(\bar{x}) = \{f: \Theta \to X \mid \forall \theta, f(\theta) \notin L_\theta^*\}$. *Thus, if* $A = \Theta$, *then* $D(\bar{x}) = B^{b^*} \in \mathcal{B}$.

(*ii*) *If* $A = \Theta$ *and* $B_{K,\mathcal{C}} = D(\bar{x})$, *then the IC and IR constraints are satisfied.*

(*iii*) *If* $A = \Theta$, *then any belief state satisfying the IC and IR constraints is a subset of* $D(\bar{x})$.

(*iv*) *If* $\bar{K} = 1$, *then every* $f \in D(\bar{x})$ *is trivial.*

PROOF OF (*i*):

Let $B = \{f: \Theta \to X \mid \forall \theta, f(\theta) \notin L_\theta^*\}$. The claim is that $D(\bar{x}) = B$.

To establish $D(\bar{x}) \subseteq B$, let $f \in D(\bar{x})$ and $\theta \in \Theta$. By definition, there is a θ^* such that $L_\theta^* = L_\theta(Y)$ where $Y = L_{\theta^*}(\bar{x}_{\theta^*})$. Since L_θ^* is a strict lower contour of $\succsim_\theta$, there is an $x^* \in X$ such that $L_\theta^* = L_\theta(x^*)$. Then $L_{\theta^*}(\bar{x}_{\theta^*}) \supseteq L_\theta(x^*)$, so that $f(\theta) \succsim_\theta x^*$ by IR-Dominance. Since $x^* \succ_\theta x$ for all $x \in L_\theta(x^*) = L_\theta^*$, it follows that $f(\theta) \notin L_\theta^*$.

For the converse inclusion, suppose $f \in B$, $\theta \in \Theta$, and $L_{\theta'}(\bar{x}_{\theta'}) \supseteq L_\theta(x)$. Since $f \in B$, we have $f(\theta) \notin L_\theta^*$. Therefore, $f(\theta) \succ_\theta x'$ for all $x' \in L_\theta^*$ because L_θ^* is a lower contour of $\succsim_\theta$. In particular, $f(\theta) \succsim_\theta x$ because $L_\theta(x) \subseteq L_\theta\big(L_\theta(x)\big) \subseteq L_\theta\big(L_{\theta'}(\bar{x}_{\theta'})\big) \subseteq L_\theta^*$ (the second inclusion holds because $L_\theta(x) \subseteq L_{\theta'}(\bar{x}_{\theta'})$ and $L_\theta(Y) \subseteq L_\theta(Y')$ whenever $Y \subseteq Y'$). Thus, $f \in D(\bar{x})$. ∎

PROOF OF (*ii*):

By (i), we may represent $D(\bar{x})$ by the set $B = \{f: \Theta \to X \mid \forall \theta, f(\theta) \notin L_\theta^*\}$. Clearly, this set satisfies the IR condition. For the IC condition, suppose toward a contradiction that some type θ strictly prefers to misreport as $\theta' \neq \theta$ under beliefs B. By the maxmin criterion, this implies $L_\theta^* \subsetneq L_\theta(L_{\theta'}^*)$; that is, $L_{\theta'}^*$ contains

a strictly larger lower contour set of $\succsim_\theta$ than L^*_θ. Now, there is a θ^* such that $L^*_{\theta'} = L_{\theta'}\big(L_{\theta^*}(\bar{x}_{\theta^*})\big)$. Then $L^*_{\theta'} \subseteq L_{\theta^*}(\bar{x}_{\theta^*})$, which implies $L_\theta(L^*_{\theta'}) \subseteq L_\theta\big(L_{\theta^*}(\bar{x}_{\theta^*})\big)$. But then $L^*_\theta \subsetneq L_\theta(L^*_{\theta'}) \subseteq L_\theta\big(L_{\theta^*}(\bar{x}_{\theta^*})\big)$. This contradicts the fact that L^*_θ is the largest set of the form $L_\theta\big(L_{\theta''}(\bar{x}_{\theta''})\big)$ among all $\theta'' \in \Theta$. Thus, $D(\bar{x})$ satisfies the IC condition as well. ∎

PROOF OF (*iii*):

Suppose B satisfies the IC and IR constraints. Let b denote the associated correspondence, and suppose toward a contradiction that there exists $(\theta, x) \in \Theta \times X$ such that $x \in b(\theta)$ but $x \notin b^*(\theta)$. Then $x \in L^*_\theta$ (because $x \notin b^*(\theta) = X \backslash L^*_\theta$ by part (*i*)) and $x \notin L_\theta(\bar{x}_\theta)$ (because $x \in b(\theta) \subseteq X \backslash L_\theta(\bar{x}_\theta)$ by IR). By definition of L^*_θ, there exists θ^* such that $L^*_\theta = L_\theta\big(L_{\theta^*}(\bar{x}_{\theta^*})\big)$. We must have $\theta^* \neq \theta$; otherwise, $L^*_\theta = L_\theta(\bar{x}_\theta)$, contradicting the fact that $x \in L^*_\theta \backslash L_\theta(\bar{x}_\theta)$.

Next, observe that if $y \in L_{\theta^*}(\bar{x}_{\theta^*})$, then $y \notin b(\theta^*)$ by the IR constraint for type θ^*. Then $z \notin b(\theta^*)$ for all $z \in L_\theta\big(L_{\theta^*}(\bar{x}_{\theta^*})\big) \subseteq L_{\theta^*}(\bar{x}_{\theta^*})$. Thus, under beliefs b, type θ expects (by the worst-case criterion) an outcome strictly better than x from reporting as type θ^*, because $x \in L^*_\theta = L_\theta\big(L_{\theta^*}(\bar{x}_{\theta^*})\big)$ and no element of $L_\theta\big(L_{\theta^*}(\bar{x}_{\theta^*})\big)$ (hence, no element $y \precsim_\theta x$) is a member of $b(\theta^*)$. This contradicts the fact that b satisfies the IC and IR constraints. ∎

PROOF OF (*iv*):

If $\min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b^*(\theta')| = 1$, then there is a θ^* such that $|b^*(\theta)| = 1$ for all $\theta \neq \theta^*$. By (i), for each $\theta \neq \theta^*$, there is a strict lower contour set L_θ such that $b^*(\theta) = X \backslash L_\theta$. Thus, the fact that $|b^*(\theta)| = 1$ implies that the sole member x_θ of $b^*(\theta)$ is an optimal outcome for type θ: $x_\theta \succsim_\theta x$ for all $x \in X$. Hence, any selection g from b^* has the property that $x_\theta = g(\theta) \succsim_\theta g(\theta')$ and $g(\theta) \succsim_\theta \bar{x}_\theta$ for all $\theta \neq \theta^*$ and all $\theta' \in \Theta$.

Now consider type θ^*. Since b^* satisfies the IC and IR constraints (claim (ii)) and $g(\theta) = g'(\theta)$ for all $\theta \neq \theta^*$ and $g, g' \in D(\bar{x})$, we have $\min_{x \in b^*(\theta^*)} u_{\theta^*}(x) \geq u_{\theta^*}\big(g(\theta)\big)$ for all $\theta \in \Theta$ and $g \in D(\bar{x})$. Thus, for every $g \in D(\bar{x})$, we have $g(\theta^*) \succsim_{\theta^*} g(\theta)$ for all θ and $g(\theta^*) \succsim_{\theta^*} \bar{x}_{\theta^*}$. Hence, every $g \in D(\bar{x})$ is trivial. ∎

LEMMA A.7: *For all K, either $B_{K,\mathcal{C}_f} = D(\bar{x})$ or $B_{K,\mathcal{C}_f} = \{f\}$. In particular, $B_{K,\mathcal{C}_f} = D(\bar{x})$ for all $K < \bar{K}$, and $B_{K,\mathcal{C}_f} = \{f\}$ for all $K \geq \bar{K}$.*

PROOF:

Let $K \geq 1$. Observe that $D(\bar{x}) \in \mathcal{B}$ is K-reachable under $\mathcal{C}_f$ because every $C \in \mathcal{C}_f$ is a subset of $D(\bar{x})$ and, hence, $G \overset{\{C\}}{\to} D(\bar{x})$ is K-valid. Thus, $B_{K,\mathcal{C}_f} \subseteq D(\bar{x})$ by Lemma A.5. To prove the first claim of this lemma, it will suffice to show that if some $B \in \mathcal{B}$ such that $B \subsetneq D(\bar{x})$ is K-reachable, then $B_{K,\mathcal{C}_f} = \{f\}$.

If some $B \subsetneq D(\bar{x})$ is K-reachable, then (by Lemma A.5) B is K-reachable from $D(\bar{x})$ because $D(\bar{x})$ is K-reachable. So, there exist $\mathcal{C}^1, \ldots, \mathcal{C}^n \subseteq \mathcal{C}_f$ and $B^1, \ldots, B^n \in \mathcal{B}$ such that $B^i \subseteq D(\bar{x})$ for all $i \geq 1$ and

$$G = B^0 \overset{\mathcal{C}^1}{\to} B^1 \overset{\mathcal{C}^2}{\to} \cdots \overset{\mathcal{C}^n}{\to} B^n = B$$

is a sequence of K-valid transitions. Let i^* be the smallest i such that $B^i \subsetneq D(\bar{x})$ and let $B' = B^{i^*}$.

Letting $\mathcal{C}' = \mathcal{C}^{i^*}$, it follows from our choice of i^* that $D(\bar{x}) \xrightarrow{\mathcal{C}'} B'$ is K-valid. Moreover, since $B' \subsetneq D(\bar{x})$, there exists $(\theta, x) \in \Theta \times X$ such that $x \in b^*(\theta)$ but $x \notin b^{B'}(\theta)$. That is, every $g \in B'$ satisfies $g(\theta) \neq x$. Hence, $\mathcal{C}'$ is of the form $\mathcal{C}' = \{D(\bar{x}) \setminus \{g'\} : g' \in E\}$ for some E containing every $g' \in D(\bar{x})$ such that $g'(\theta) = x$. Thus, since $|\mathcal{C}'| \leq K$,

$$|\{g \in D(\bar{x}) : g(\theta) = x\}| \leq K. \tag{4}$$

Note that (4) holds for every choice of $x \in b^*(\theta)$ such that $x \neq g_{\mathcal{C}_f}(\theta)$ because $|\{g \in D(\bar{x}) : g(\theta) = x\}| = \prod_{\theta' \neq \theta} |b^*(\theta')|$, which does not depend on x.

So, suppose $b^*(\theta) \setminus \{g_{\mathcal{C}_f}(\theta)\} = \{x_1, \ldots, x_m\}$. For each x_i, let

$$\mathcal{C}^{(\theta, x_i)} := \{D(\bar{x}) \setminus \{g\} : g \in D(\bar{x}) \text{ and } g(\theta) = x_i\}.$$

Clearly $\mathcal{C}^{(\theta, x_i)} \subseteq \mathcal{C}_f$. Moreover,

$$D(\bar{x}) \xrightarrow{\mathcal{C}^{(\theta, x_1)}} \hat{B}^1 \xrightarrow{\mathcal{C}^{(\theta, x_2)}} \cdots \xrightarrow{\mathcal{C}^{(\theta, x_m)}} \hat{B}^m$$

is a sequence of K-valid transitions where, for every $i = 1, \ldots, m$, $\hat{B}^i$ satisfies $b^{\hat{B}^i}(\theta) = b^*(\theta) \setminus \{x_1, \ldots, x_i\}$. The transitions are K-valid because $|\mathcal{C}^{(\theta, x_i)}| = |\{g \in D(\bar{x}) : g(\theta) = x\}|$, which does not exceed K by (4).

Notice that every $g \in \hat{B}^m$ satisfies $g(\theta) = g_{\mathcal{C}_f}(\theta)$. In other words, the fact that some $x \in b^*(\theta)$ ($x \neq g_{\mathcal{C}_f}(\theta)$) is eliminated in state B' implies the agent is, in fact, sophisticated enough to pin down $g_{\mathcal{C}_f}(\theta)$ after a series of K-valid transitions.

For each nonempty $\Theta' \subseteq \Theta$, let $B_{-\Theta'} := \{g \in D(\bar{x}) : \forall \theta' \in \Theta', g(\theta') = g_{\mathcal{C}_f}(\theta')\}$. Clearly $B_{-\Theta'} \in \mathcal{B}$, and the argument above shows that $B_{-\{\theta\}}$ is K-reachable. To complete the proof, I show that if some $B_{-\Theta'}$ with $\theta \in \Theta'$ is K-reachable, then so is $B_{-\Theta' \cup \{\theta'\}}$ for any $\theta' \in \Theta \setminus \Theta'$. This implies that an agent who can eliminate at least one point from the correspondence b^* will actually deduce $g_{\mathcal{C}}$. Since ability $\bar{K} = \min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b^*(\theta')|$ is required to eliminate such a point, the second claim of the lemma follows immediately.

Let $\theta' \in \Theta \setminus \Theta'$. If $x' \in b^*(\theta')$ and $x' \neq g_{\mathcal{C}_f}(\theta')$, then

$$\begin{aligned} |\{g \in B_{-\Theta'} : g(\theta') = x'\}| &= \prod_{\hat{\theta} \in \Theta \setminus (\Theta' \cup \theta')} |b^*(\hat{\theta})| \\ &\leq \prod_{\hat{\theta} \in \Theta \setminus \theta} |b^*(\hat{\theta})| \qquad \text{since } \theta \in \Theta' \\ &\leq K \qquad \text{by (4).} \end{aligned}$$

It follows that $|\hat{\mathcal{C}}^{(\theta', x')}| \leq K$ for all such x', where

$$\hat{\mathcal{C}}^{(\theta', x')} := \{D(\bar{x}) \setminus \{g\} : g \in B_{-\Theta'} \text{ and } g(\theta') = x'\} \subseteq \mathcal{C}_f.$$

Hence, if $b^*(\theta')\backslash\{g_{\mathcal{C}_f}(\theta')\} = \{x'_1, \ldots, x'_\ell\}$, then

$$B_{-\Theta'} \overset{\hat{\mathcal{C}}^{(\theta',x'_1)}}{\to} \hat{B}^1_{-\Theta'} \overset{\hat{\mathcal{C}}^{(\theta',x'_2)}}{\to} \cdots \overset{\hat{\mathcal{C}}^{(\theta',x'_\ell)}}{\to} \hat{B}^\ell_{-\Theta'}$$

is a sequence of K-valid transitions where $B^i_{-\Theta'} \in \mathcal{B}$ satisfies $b^{B^i_{-\Theta'}}(\theta') = b^*(\theta')\backslash\{x'_1, \ldots, x'_i\}$, so that $B^\ell_{-\Theta'} = B_{-\Theta'\cup\{\theta'\}}$ is K-reachable. ∎

LEMMA A.8: *If a function is implementable, then it is IR-Dominant.*

PROOF:

Suppose a contract, $\mathcal{C}$, K-implements f. Then there is an action profile $(a_\theta)_{\theta\in\Theta}$ such that $U_\theta(a_\theta, K, \mathcal{C}) \geq U_\theta(a', K, \mathcal{C})$ for all $\theta \in \Theta$ and $a' \in A$. Define $b : \Theta \rightrightarrows X$ by $b(\theta) := b_{K,\mathcal{C}}(a_\theta)$. Then, by construction, b is incentive-compatible. By part (iii) of Lemma A.6, then, $B^b \subseteq D(\bar{x})$, where $B^b = \{g : \Theta \to X \,|\, \forall\theta,\ g(\theta) \in b(\theta)\}$. Since $f(\theta) = g_{\mathcal{C}}(\theta) \in b_{K,\mathcal{C}}(a_\theta) = b(\theta)$ for all θ, it follows that f is IR-Dominant. ∎

PROOF OF THEOREM 1:

Suppose f is implementable. By Lemma A.8, f is IR-Dominant. Therefore, the contract $\mathcal{C}_f$ is well defined and satisfies $g_{\mathcal{C}_f} = f$.

Consider an agent of ability $K = 1$. If $\bar{K} > 1$, then (by Lemma A.7) we have $B_{1,\mathcal{C}_f} = D(\bar{x})$. By part (ii) of Lemma A.6, these beliefs satisfy the IR and IC constraints. If instead $\bar{K} = 1$, the agent arrives at belief state $B_{1,\mathcal{C}_f} = \{f\}$. By part (iv) of Lemma A.6, f is trivial. Thus, in each case, $B_{1,\mathcal{C}_f}$ satisfies IR and IC. Consequently, f is 1-implementable by $\mathcal{C}_f$ (a direct contract). ∎

PROOF OF THEOREM 2:

By Lemma A.8, IR-Dominance is a necessary condition for implementability. To prove the remaining statements of the theorem, an additional lemma is required.

LEMMA A.9: *If a contract, $\mathcal{C}$, directly K-implements a function f, then $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$.*

PROOF:

Clearly, state $D(\bar{x}) \in \mathcal{B}$ is K-reachable under $\mathcal{C}_f$ for all K (take $\mathcal{C}' = \{C\}$ for any $C \in \mathcal{C}_f$ to get that $G \overset{\mathcal{C}'}{\to} D(\bar{x})$ is K-valid). By part (iii) of Lemma A.6, we have $B_{K,\mathcal{C}} \subseteq D(\bar{x})$. Thus, $D(\bar{x})$ is K-reachable under $\mathcal{C}$ as well. I prove that if $B \overset{\mathcal{C}'}{\to} B'$ is K-valid for some $B, B' \subseteq D(\bar{x})$ and $\mathcal{C}' \subseteq \mathcal{C}_f$, then there is a $\hat{\mathcal{C}} \subseteq \mathcal{C}$ such that $B \overset{\hat{\mathcal{C}}}{\to} B'$ is K-valid. This implies that every state that is K-reachable from $D(\bar{x})$ under $\mathcal{C}_f$ is also K-reachable from $D(\bar{x})$ under $\mathcal{C}$. Then $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$ by Lemma A.5.

So, suppose $B \overset{\mathcal{C}'}{\to} B'$ is K-valid for some $B, B' \subseteq D(\bar{x})$ and $\mathcal{C}' \subseteq \mathcal{C}_f$. Then there exists $g_1, \ldots, g_n \in D(\bar{x})$ (where $n \leq K$) such that $\mathcal{C}' = \{D(\bar{x})\backslash\{g_i\} : i = 1, \ldots, n\} \subseteq \mathcal{C}_f$ and

$$B \cap \left(\bigcap_{C\in\mathcal{C}'} C\right) \subseteq B'. \tag{5}$$

Note that $g_{\mathcal{C}} = f \neq g_i$ for all i. Thus, for each $i = 1, \ldots, n$ there exists $C^i \in \mathcal{C}$ such that $g_i \notin C^i$. Take $\hat{\mathcal{C}} = \{C^i : i = 1, \ldots, n\}$ and observe that $B \cap C^i \subseteq D(\bar{x}) \backslash \{g_i\}$. Then,

$$\begin{aligned} B \cap \left(\bigcap_{C \in \hat{\mathcal{C}}} C \right) &= B \cap \left(\bigcap_{i=1}^{n} \left(B \cap C^i \right) \right) \\ &\subseteq B \cap \left(\bigcap_{i=1}^{n} \left(D(\bar{x}) \backslash \{g_i\} \right) \right) \\ &= B \cap \left(\bigcap_{C \in \mathcal{C}'} C \right). \end{aligned}$$

Combined with (5), it follows that

$$B \cap \left(\bigcap_{C \in \hat{\mathcal{C}}} C \right) \subseteq B'$$

so that $B \xrightarrow{\hat{\mathcal{C}}} B'$ is K-valid. ∎

To complete the proof of Theorem 2, suppose f is a nontrivial, IR-Dominant function. We may assume $\bar{K} > 1$ (otherwise, by part (iv) of Lemma A.6, f is trivial). By Lemma A.7, $B_{K,\mathcal{C}_f} = D(\bar{x})$ for all $1 \leq K < \bar{K}$. This belief state is incentive-compatible by part (ii) of Lemma A.6, and therefore $\mathcal{C}_f$ K-implements f for all $K < \bar{K}$. Thus, for nontrivial, IR-Dominant functions, $K < \bar{K}$ is a sufficient condition for K-implementability.

To see that $K < \bar{K}$ is also a necessary condition, suppose a contract $\mathcal{C}$ K-implements f. By Lemma A.9, we have $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$. If $K \geq \bar{K}$, then (by Lemma A.7) $B_{K,\mathcal{C}_f} = \{f\}$, forcing $B_{K,\mathcal{C}} = \{f\}$. This contradicts the fact that $\mathcal{C}$ K-implements the (nontrivial) function f. Thus, a nontrivial, IR-Dominant function is (directly) K-implementable if and only if $K < \bar{K}$, and $\mathcal{C}_f$ achieves K-implementation for all such K. ∎

C. *Proof of Proposition 1*

Observe that if $\bar{x}'_\theta \succsim_\theta \bar{x}_\theta$ for all θ, then $L_\theta(\bar{x}'_\theta) \supseteq L_\theta(\bar{x}_\theta)$ for all θ and, hence, L^*_θ is larger under $\bar{x}'$ than $\bar{x}$ (for all θ). By Lemma A.6, it follows that $D(\bar{x}') \subseteq D(\bar{x})$. Thus, if b' and b are the belief correspondences associated with $D(\bar{x}')$ and $D(\bar{x})$, respectively, then $b'(\theta) \subseteq b(\theta)$. Therefore,

$$\bar{K}(\bar{x}') = \min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b'(\theta')| \leq \min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b(\theta')| = \bar{K}(\bar{x}). \blacksquare$$

D. *Proof of Proposition 2*

As in Section IIIC, write $A = \Theta \cup A'$ (a disjoint union). By construction, $D^A(\bar{x})$ is represented by a correspondence $b : A \rightrightarrows X$ such that $b(a) = X$ for all $a \in A'$ and $b(\theta) = b^*(\theta)$ for $\theta \in \Theta$, where b^* is the correspondence associated with $D(\bar{x})$. By the maxmin criterion, we may restrict attention to actions in the set Θ. By Lemma A.6, then, beliefs b make truthful reporting optimal for all types.

Since each $C \in \mathcal{C}_f^A$ is a subset of $D^A(\bar{x})$, the state $D^A(\bar{x})$ is K-reachable for all K. By an argument similar to that of Lemma A.7, then, we see that $B_{K,\mathcal{C}_f^A} = D^A(\bar{x})$ for all $K < \bar{K}^A$. Thus, $\mathcal{C}_f^A$ K-implements f for all $K < \bar{K}^A$. ∎

E. *Proof of Proposition 3*

Suppose the agent has α-maxmin preferences and let $f \in D(\bar{x})$. By Lemma A.7, $B_{K,\mathcal{C}_f} = D(\bar{x})$ for all $K < \bar{K}$. By Lemma A.6, beliefs $D(\bar{x})$ are represented by a correspondence b^* where $b^*(\theta) = X \backslash L_\theta^*$, where L_θ^* is a particular strict lower-contour set of $\succsim_\theta$. Let $\theta \in \Theta$. Since $D(\bar{x})$ is incentive-compatible (part (ii) of Lemma A.6), we have $\min_{x \in b^*(\theta)} u_\theta(x) \geq \min_{x \in b^*(\theta')} u_\theta(x)$ for all θ', and also $\min_{x \in b^*(\theta)} u_\theta(x) \geq u_\theta(\bar{x})$. Since $b^*(\theta)$ is formed by removing a strict lower-contour of $\succsim_\theta$ from X, it follows that every $\succsim_\theta$-maximal outcome is a member of $b^*(\theta)$ (that is, type θ believes his most-preferred outcome(s) are possible consequences of truthful reporting under beliefs b^*). Therefore, for all $\theta, \theta' \in \Theta$, $\max_{x \in b^*(\theta)} u_\theta(x) \geq \max_{x \in b^*(\theta')} u_\theta(x)$. Thus, under beliefs b^*, $U_\theta^\alpha(\theta) \geq U_\theta^\alpha(\theta')$ and $U_\theta^\alpha(\theta) \geq u_\theta(\bar{x})$ for all $\theta, \theta' \in \Theta$, so that $\mathcal{C}_f$ implements f for all $K < \bar{K}$. ∎

F. *Proof of Proposition 4*

If f is c-implementable by a contract $\mathcal{C}$, then K^* induces incentive-compatible beliefs $B_{K^*,\mathcal{C}}$ under $\mathcal{C}$: in other words, $\mathcal{C}$ must K^*-implement f. Therefore, by part (iii) of Lemma A.6, $f \in B \subseteq D(\bar{x})$, so that f is IR-Dominant.

Now let $f \in D(\bar{x})$ be nontrivial. By Lemma A.7, $B_{K,\mathcal{C}_f} = D(\bar{x})$ for $K < \bar{K}$ and $B_{K,\mathcal{C}_f} = \{f\}$ for $K \geq \bar{K}$. Since c is strictly increasing in K, this means an agent of type θ chooses either $K^* = 1$ or $K^* = \bar{K}$. If $u_\theta^*(f) - c(\bar{K}) < \min_{g \in D(\bar{x})} u_\theta(g(\theta))$, then the payoff from acquiring ability $\bar{K}$ and perfectly manipulating the mechanism does not exceed the (maxmin) payoff of acquiring $K = 1$ and reporting truthfully under beliefs $B_{1,\mathcal{C}_f} = D(\bar{x})$. Thus, if $\max_{\theta \in \Theta} u_\theta^*(f) - \min_{g \in D(\bar{x})} u_\theta(g(\theta)) < c(\bar{K})$, each type θ chooses $K^* = 1$ and $\mathcal{C}_f$ c-implements f. ∎

G. *Proof of Proposition 5*

Suppose $T = (K, \hat{\mathcal{B}})$ and that $\mathcal{C}$ T-implements a function f. Let $B_{T,\mathcal{C}}^*$ denote the effective belief for T given $\mathcal{C}$, and let $\mathcal{C}_{T,f} := \{B_{T,\mathcal{C}}^* \backslash \{g\} : g \in B_{T,\mathcal{C}}^* \backslash \{f\}\}$. Let $T' = (K', \hat{\mathcal{B}}') \leq T$. We must have $f \in B_{T,\mathcal{C}}^* \subseteq D(\bar{x})$ since $B_{T,\mathcal{C}}^*$ is incentive-compatible (part (iii) of Lemma A.6). Thus, f is IR-Dominant.

LEMMA A.10: *If there exist $B, B' \in \hat{\mathcal{B}}'$ such that $B, B' \subseteq B_{T,\mathcal{C}}^*$ and $\mathcal{C}' \subseteq \mathcal{C}_{T,f}$ such that the transition $B \overset{\mathcal{C}'}{\to} B'$ is T'-valid, then there exists $\mathcal{C}'' \subseteq \mathcal{C}$ such that $B \overset{\mathcal{C}''}{\to} B'$ is T-valid.*

PROOF:

Suppose $B, B' \in \hat{\mathcal{B}}'$ (hence, $B, B' \in \hat{\mathcal{B}}$) and that $B \overset{\mathcal{C}'}{\to} B'$ is T'-valid for some $\mathcal{C}' \subseteq \mathcal{C}_{T,f}$. Then $|\mathcal{C}'| \leq K' \leq K$ and $B \cap \left(\bigcap_{C' \in \mathcal{C}'} C'\right) \subseteq B'$. Every clause $C' \in \mathcal{C}'$ is of the form $B_{T,\mathcal{C}}^* \backslash \{g\}$, where $g \in B_{T,\mathcal{C}}^* \backslash \{f\}$. Thus, $\bigcap_{C' \in \mathcal{C}'} C' = B_{T,\mathcal{C}}^* \backslash \{g_1, \ldots, g_{\hat{K}}\}$ where $\hat{K} \leq K'$ and $f \neq g_i \in B_{T,\mathcal{C}}^*$ for all

$i = 1, \ldots, \hat{K}$. For each i, choose $C_i \in \mathcal{C}$ such that $g_i \notin C_i$; such clauses exist because $g_{\mathcal{C}} = f \neq g_i$. Note that $B^*_{T,\mathcal{C}} \cap C_i \neq \varnothing$ because $f \in B^*_{T,\mathcal{C}} \cap C_i$. Finally, let $\mathcal{C}'' = \{C_1, \ldots, C_{\hat{K}}\}$ and observe that $|\mathcal{C}''| \leq \hat{K}$ (with strict inequality if $C_i = C_j$ for some $i \neq j$). Then $B \cap \bigcap_{C'' \in \mathcal{C}''} C'' \subseteq B \cap \bigcap_{C' \in \mathcal{C}'} C' \subseteq B'$. ∎

To complete the proof of Proposition 5, observe that $B^*_{T,\mathcal{C}}$ is T''-reachable under $\mathcal{C}_{T,f}$ for all types T''. Thus, for type T', the induced belief under contract $\mathcal{C}_{T,f}$ is the intersection of all states in $\hat{\mathcal{B}}'$ that are T'-reachable from $B^*_{T,\mathcal{C}}$. Let $B^*_{T',\mathcal{C}_{T,f}}$ denote the effective induced belief for T' under $\mathcal{C}_{T,f}$. By the preceding discussion, $B^*_{T',\mathcal{C}_{T,f}} \subseteq B^*_{T,\mathcal{C}}$. Suppose toward a contradiction that $B^*_{T',\mathcal{C}_{T,f}} \subsetneq B^*_{T,\mathcal{C}}$. Then there is a sequence

$$B^*_{T,\mathcal{C}} = B^0 \overset{\mathcal{C}^1}{\to} \hat{B}^1 \overset{\mathcal{C}^2}{\to} \hat{B}^2 \overset{\mathcal{C}^3}{\to} \cdots \overset{\mathcal{C}^n}{\to} B^n = B \subseteq B^*_{T',\mathcal{C}_{T,f}} \subsetneq B^*_{T,\mathcal{C}}$$

of T'-valid transitions where $\mathcal{C}^i \subseteq \mathcal{C}_{T,f}$ for all i. By Lemma A.10, each transition $B^{i-1} \overset{\mathcal{C}^i}{\to} B^i$ can be replaced by a T-valid transition from B^{i-1} to B^i under contract $\mathcal{C}$. This implies that a state $B \subseteq B^*_{T',\mathcal{C}_{T,f}} \subsetneq B^*_{T,\mathcal{C}}$ is T-reachable under $\mathcal{C}$, contradicting the fact that $B^*_{T,\mathcal{C}}$ is the effective induced belief for T under $\mathcal{C}$. Thus, $B^*_{T',\mathcal{C}_{T,f}} = B^*_{T,\mathcal{C}}$, so that $\mathcal{C}_{T,f}$ T'-implements f. ∎

REFERENCES

Alaoui, Larbi, and Antonio Penta. 2016a. "Endogenous Depth of Reasoning." *Review of Economic Studies* 83 (4): 1297–1333.

Alaoui, Larbi, and Antonio Penta. 2016b. "Cost-Benefit Analysis in Reasoning." Unpublished.

Battigalli, Pierpaolo, and Giovanni Maggi. 2002. "Rigidity, Discretion, and the Costs of Writing Contracts." *American Economic Review* 92 (4): 798–817.

Bodoh-Creed, Aaron L. 2012. "Ambiguous Beliefs and Mechanism Design." *Games and Economic Behavior* 75 (2): 518–37.

Bose, Subir, and Arup Daripa. 2009. "A Dynamic Mechanism and Surplus Extraction under Ambiguity." *Journal of Economic Theory* 144 (5): 2084–2114.

Bose, Subir, Emre Ozdenoren, and Andreas Pape. 2006. "Optimal Auctions with Ambiguity." *Theoretical Economics* 1 (4): 411–38.

Bose, Subir, and Ludovic Renou. 2014. "Mechanism Design with Ambiguous Communication Devices." *Econometrica* 82 (5): 1853–72.

de Clippel, Geoffroy. 2014. "Behavioral Implementation." *American Economic Review* 104 (10): 2975–3002.

de Clippel, Geoffroy, Rene Saran, and Roberto Serrano. 2018. "Level-*k* Mechanism Design." *Review of Economic Studies* 86 (3): 1207–27.

di Tillio, Alfredo, Nenad Kos, and Matthias Messner. 2016. "The Design of Ambiguous Mechanisms." *Review of Economic Studies* 84 (1): 237–76.

Eliaz, Kfir. 2002. "Fault Tolerant Implementation." *Review of Economic Studies* 69 (3): 589–610.

Ellsberg, Daniel. 1961. "Risk, Ambiguity, and the Savage Axioms." *Quarterly Journal of Economics* 75 (4): 643–69.

Gilboa, Itzhak, and David Schmeidler. 1989. "Maxmin Expected Utility with Non-Unique Prior." *Journal of Mathematical Economics* 18 (2): 141–53.

Glazer, Jacob, and Ariel Rubinstein. 2012. "A Model of Persuasion with Boundedly Rational Agents." *Journal of Political Economy* 120 (6): 1057–82.

Glazer, Jacob, and Ariel Rubinstein. 2014. "Complex Questionnaires." *Econometrica* 82 (4): 1529–41.

Hurwicz, Leonid. 1951. "Some Specification Problems and Applications to Econometrics Models." *Econometrica* 19 (3): 343–44.

Kneeland, Terri. 2018. "Mechanism Design with Level-*k* Types: Theory and an Application to Bilateral Trade." Unpublished.

Korpela, Ville. 2012. "Implementation without Rationality Assumptions." *Theory and Decision* 72 (2): 189–203.

Kőszegi, Botond. 2014. "Behavioral Contract Theory." *Journal of Economic Literature* 52 (4): 1075–1118.

Lipman, Barton L. 1999. "Decision Theory without Logical Omniscience: Toward an Axiomatic Framework for Bounded Rationality." *Review of Economic Studies* 66 (2): 339–61.

Li, Shengwu. 2017. "Obviously Strategy-Proof Mechanisms." *American Economic Review* 107 (11): 3257–87.

Salant, Yuval, and Ariel Rubinstein. 2008. "(A, f): Choice with Frames." *Review of Economic Studies* 75 (4): 1287–96.

Salant, Yuval, and Ron Siegel. 2018. "Contracts with Framing." *American Economic Journal: Microeconomics* 10 (3): 315–46.

Stahl, Dale O., and Paul W. Wilson. 1994. "Experimental Evidence on Players' Models of Other Players." *Journal of Economic Behavior & Organization* 25 (3): 309–27.

Stahl, Dale O., and Paul W. Wilson. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior* 10 (1): 218–54.

Wolitzky, Alexander. 2016. "Mechanism Design with Maxmin Agents: Theory and an Application to Bilateral Trade." *Theoretical Economics* 11 (3): 971–1004.