A Model of Complex Contracts

Alexander M. Jakobsen*†

Abstract

I study a behavioral mechanism design problem involving a principal and a single agent. The principal seeks to implement a function mapping agent types to outcomes and must commit to a mechanism. Only a small class of functions are implementable if the agent is fully rational. I introduce a model of bounded rationality where the agent has limited ability to combine different pieces of information and to retain new facts derived in the process. The agent transitions among coarse 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, expanding the set of implementable functions. I characterize the set of implementable functions and show that, without loss of generality, the principal selects a contract achieving implementation for the widest possible range of K. The optimal contract is also robust to several variations on the cognitive procedure.

*This version: February 18, 2019

†Department of Economics, University of Calgary; alexander.jakobsen@ucalgary.ca. 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 indebted to Faruk Gul and Wolfgang Pesendorfer for their guidance and support. Thanks also to Dilip Abreu, Roland Benabou, 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.

1 Introduction

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 whether the underlying game induces incentives for truthful reporting. In this paper, I study a mechanism design problem where 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 and imposing large costs of compliance on economic agents. 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 constrained agents? How, and to what degree, can designers accommodate heterogeneous cognitive procedures or abilities?

The starting point of my analysis 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 to the true mechanism. Thus, in my framework, complexity stems from the way mechanisms are expressed—two different presentations (contracts) of the same mechanism may differ in the cognitive resources required to identify the mechanism. My concept of bounded rationality rests on basic principles of framing and procedural reasoning and, as explained in section 5, 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.

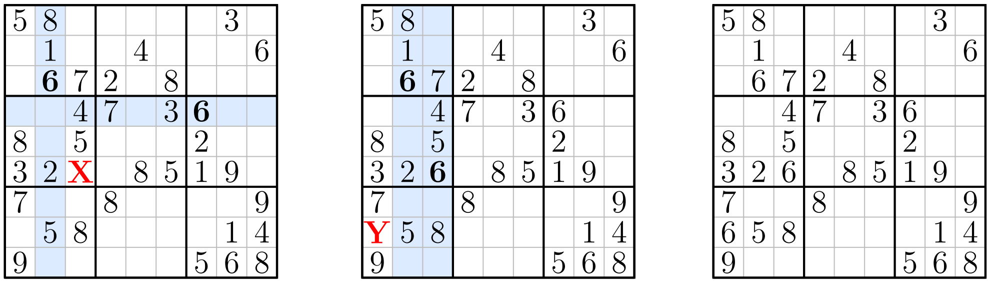
To illustrate the procedure, consider the game Sudoku. In this game, a player is presented with a 9×9 table. Some cells are initially filled with entries from the set $D = \{1, 2, \ldots, 9\}$ and the player must deduce the entries for the remaining cells. The rules are that each digit $d \in D$ must 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 (Figure 1a). Combined with rule (iii), this implies X (position (6,3)) must be 6. He updates the configuration to reflect this (Figure 1b). 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 (Figure 1c). 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

[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.



(a) Player deduces X = 6

(b) . . . then Y = 6

(c) New configuration

Figure 1: A possible sequence of deductions in Sudoku

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 X = 6, and three again to deduce 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.

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 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 objective functions are implementable under full rationality: for 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.

My first result, Proposition 1, establishes that the set of implementable functions is fully characterized by a simple *IR-Dominance* condition. Suppose types θ and θ' agree 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. Proposition 1 states that a function is implementable (for some K) if and only if the function is IR-Dominant, but makes no claim about the structure of implementing contracts or the values of K that are supported.

The main result, Proposition 2, establishes that if a function f is implementable for some K, then a particular contract (the *complex contract* for f) implements f for all $K' \leq K$. It follows that the principal can do no better than to choose the complex contract for f, and that she need not cater contracts to particular values of K. The contract is designed according to a simple principle: maximize the difficulty of performing transitions subject to the constraint that each clause makes truthful reporting appear optimal. Consequently, the complex contract satisfies many robustness properties. For example, one can introduce randomness, impatience, or costs and benefits of reasoning (thereby endogenizing K) without severely undermining the effectiveness of the contract. I discuss these robustness properties in section 3.4, and study a set of variations and extensions (including richer belief states) in section 4. Section 3.3 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 5). This paper is part of the emerging literature on behavioral mechanism design. Most of this literature involves agents who understand game forms and mechanisms but exhibit non-standard choice behavior (reference dependence, present bias, etc) or limited strategic reasoning (eg, level-k reasoning in games). In contrast, I focus on how cognitive limitations influence the agent's understanding of the mechanism itself.[2] The cognitive limitation is modeled as a sensitivity to the way mechanisms are framed, and is quite distinct from imperfect strategic reasoning (eg, 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), who study 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 focus on persuasion, while I consider a general implementation problem with type-dependent preferences and outside options. Because of this, our models yield rather different insights. I elaborate on this, as well as other related literature, in section 5.

2 Model

2.1 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 $\overline{x}_\theta \in X$. Let $u_\theta : X \to \mathbb{R}$ be a utility function representing $\succsim_\theta$ and $\overline{x} := (\overline{x}_\theta)_{\theta \in \Theta}$ denote the full profile of outside options.

[2] 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.

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 $g(a)$. If the agent chooses not to participate in the mechanism, he consumes his outside option instead.

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". There are of course many different ways of representing a set C in formal or natural language, and this is important for the interpretation of the model (see section 2.4).

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, where designers must commit to a single mechanism. Finally, note that contracts are formalized 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.[3]

2.2 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

[3] I emphasize that, in this paper, the word "contract" has a different meaning than is usually understood in the literature. In my model, a contract is a particular way of framing or expressing a mechanism (namely, as a set of clauses).

$g_{\mathcal{C}}(a)$, the outcome actually prescribed by $\mathcal{C}$.

2.3 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 $B^b := \{g \in G \mid \forall a \; g(a) \in b(a)\}$ 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.

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

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

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 $g_{\mathcal{C}} \in B'$. 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
>
> $$G = B^0 \xrightarrow{\mathcal{C}^1} B^1 \xrightarrow{\mathcal{C}^2} B^2 \xrightarrow{\mathcal{C}^3} \dots \xrightarrow{\mathcal{C}^n} B^n = B$$
>
> 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 whenever B is K-reachable. Thus, in K-reachable states, the agent does not erroneously eliminate the true mechanism from consideration.

> **Lemma 1.**
> Let $\mathcal{C}$ be a contract and $K \geq 1$. There is a unique K-reachable state, denoted $B_{K,\mathcal{C}}$, such that $B_{K,\mathcal{C}} \subseteq B$ for all K-reachable $B \in \mathcal{B}$.

Lemma 1 states that for every contract, there exists a finest K-reachable belief state. This follows from the fact that $B \cap B'$ is K-reachable whenever B and B' are K-reachable. Thus, $B_{K,\mathcal{C}}$ is simply the intersection of all K-reachable states (see the appendix for proof). By Lemma 1, the following is well-defined:

> **Definition 3** (Induced Belief)**.**
> The *induced belief state* for an agent of ability K under contract $\mathcal{C}$ is $B_{K,\mathcal{C}}$ (Lemma 1). The associated correspondence, denoted $b_{K,\mathcal{C}}$, is the *induced belief.*

Definition 3 states that the agent arrives at the finest possible approximation to $g_{\mathcal{C}}$ given his ability K. Effectively, the agent repeatedly combines clauses and performs transitions until he gets stuck in a state where further refinement of his beliefs requires the combination of more than K clauses of $\mathcal{C}$. Lemma 1 ensures there is only one such terminal belief state, despite the many different sequences of transitions the agent may perform. The definition asserts that he reaches $B_{K,\mathcal{C}}$ but makes no claim about the sequence of transitions performed along the way.[4]

If the agent fails to deduce $g_{\mathcal{C}}$, then, from his perspective, the contract is ambiguous: there are actions $a \in A$ such that $b_{K,\mathcal{C}}(a)$ contains two or more outcomes. To close the model, an assumption regarding the agent's attitude toward such (perceived) ambiguity is required.

> **Assumption** (Ambiguity aversion)**.**
> Given a contract $\mathcal{C}$, an agent of ability K and type θ evaluates actions $a \in A$ by the formula
>
> $$\begin{aligned} U_\theta(a, K, \mathcal{C}) &:= \min_{x \in b_{K,\mathcal{C}}(a)} u_\theta(x) \\ &= \min_{g \in B_{K,\mathcal{C}}} u_\theta(g(a)) \end{aligned}$$
>
> and participates if and only if $\max_{a \in A} U_\theta(a, K, \mathcal{C}) \geq u_\theta(\overline{x}_\theta)$.

That is, he adopts a worst-case (maxmin) criterion when evaluating the set of outcomes $b_{K,\mathcal{C}}(a)$ he considers possible at actions $a \in A$. This is an extreme degree of ambiguity aversion, but many insights generated by the model hold under alternative assumptions. In fact, the main results hold even if the worst-case criterion is replaced by the Hurwicz (1951) α-criterion for any choice of α—see section 4.2.

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.

[4] In fact, Lemma 1 implies the process is *path independent*: the order of transitions does not matter because there is no chance of getting stuck in a state other than $B_{K,\mathcal{C}}$.

Example 1.
A manager is recruiting employees 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 to employees based on their type reports. 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.[5] Only by combining C_3, C_4, and C_5 (simultaneously) with B can a new belief correspondence 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.

2.4 Comments on the Agent's Procedure

The purpose of this model is to formalize basic aspects (memory and computational ability) of human problem-solving, providing an intuitive model of behavior and a framework for evaluating contract complexity. As illustrated

[5] 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.

by Example 1 and the Sudoku game in the introduction, a difficult problem is one the agent may fail to solve despite understanding all of the rules.[6] Of course, any attempt to model human behavior ignores important aspects of reality. Below, I comment on four such points.

1. Clauses as units of information. Clauses are defined as sets of mechanisms and interpreted as written or verbal statements in contracts. Since $K \geq 1$, the agent understands each clause. However, I do not assume the agent understands all the different ways a given clause might be expressed—indeed, the model is based on the idea that the agent is sensitive to the way information is framed. Rather, I assume the principal has sufficient expressive power to convey clauses as separate statements in a way the agent understands, and is willing to clarify any confusion about individual clauses.

The distinction is relevant, in part, because any contract can be read as a single statement where clauses are joined by the word "and". If, for example, φ_1 and φ_2 are statements representing clauses C_1 and C_2, then "φ_1 and φ_2" is a statement representing $C_1 \cap C_2$. Whether the agent treats "φ_1 and φ_2" as a unit or as two clauses requiring effort to combine is entirely dependent on how the principal expresses (frames) the clauses. For example, most people do not immediately replace the conjunction of clauses C_1 to C_5 in Example 1 with the (logically equivalent) statement "$L = 4$ and $H = 3$" because the clauses are framed in such a way that effort is required to combine them.

The model can be reformulated to formally distinguish clauses from their representations (frames). This allows a precise statement of the assumption regarding the principal's expressive power, but mainly adds a layer of superfluous notation to the model. As will become clear, it is in the principal's best interest to clearly express clauses as individual statements. She benefits

[6] Despite the use of states and transitions, my model has little in common with notions of complexity developed by computer scientists and formalized through the use of Turing machines. These notions revolve around worst-case run times of optimal algorithms as the input size tends to infinity. A crucial distinction is that the agent in my model can get stuck (fail to solve the problem), but such algorithms cannot—they simply take longer to run under adversarial initial inputs.

by exploiting the agent's limited memory and computational ability; poorly-framed clauses would only undermine her efforts, as would any presentation where individual clauses are not plausibly treated as units.

2. Unrestricted language. In the model, any nonempty set of mechanisms qualifies as a clause. Therefore, no set of mechanisms is too detailed or peculiar for the agent to comprehend. Again, I view this as an assumption about the expressive power of the principal: through effective use of language, she can convey any piece of information in a way the agent understands. Models of standard rational behavior implicitly make this assumption as well.

In reality, there are probably limits to what can be conveyed and understood. Even with a restricted set of permissible clauses, however, the issue of how the agent processes and combines sets of clauses would be of central importance. Focusing on this aspect (rather than the mechanics of restricted language) also reinforces a broader point: the principal *chooses* to express mechanisms as complex sets of clauses despite having the ability to communicate them clearly. Restricting the set of permissible clauses would arbitrarily introduce complexity into contracts.

3. Choice of belief states. Belief states are associated with correspondences because it seems natural for people to reason about the set of possible consequences associated with each action, rather than "correlations" between the outcomes of different actions. Hopefully, Example 1 above reinforces this position. Nonetheless, it is reasonable to wonder what would happen under alternative specifications. In section 4.3, I show how to analyze the model with richer families of belief states. The results are nearly identical to those of the baseline model.

4. Ambiguity attitude. Though not a component of the cognitive model per se, ambiguity attitude is an important determinant of the agent's behavior. It turns out that ambiguity attitude affects the set of implementable functions, but not the broader lessons derived from the model. Roughly speaking, this holds because ambiguity attitude is independent of the agent's cognitive procedure—see sections 3.4 and 4.2.

Still, there are good reasons to favor ambiguity aversion. It is a common phenomenon and, hence, a useful benchmark.[7] In the context of the implementation model, 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 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.

3 Implementation via Complex Contracts

In this section, I restrict attention to direct mechanisms (those where $A = \Theta$). The revelation principle does not hold in my model, but this restriction is almost without loss: varying the action space yields interesting comparative statics, but none of the main results are affected (see section 3.3.2). Throughout, $\overline{x} = (\overline{x}_\theta)_{\theta \in \Theta}$ denotes a fixed profile of outside options.

Definition 4 (K-Implementation)**.**
Let $K \geq 1$. A contract, $\mathcal{C}$, *K-implements* the function $f : \Theta \to X$ if, for all $\theta, \theta' \in \Theta$,

1. $U_\theta(\theta, K, \mathcal{C}) \geq U_\theta(\theta', K, \mathcal{C})$ (Incentive Compatibility),
2. $U_\theta(\theta, K, \mathcal{C}) \geq u_\theta(\overline{x}_\theta)$ (Individual Rationality), and
3. $g_{\mathcal{C}}(\theta) = f(\theta)$.

A function f is *K-implementable* if there exists a contract that

[7] Since Ellsberg (1961), many studies have replicated the finding of ambiguity aversion. Traditionally, ambiguity has been limited to the domain of probabilistic beliefs. However, recent studies such as Eliaz and Ortoleva (2015) have documented ambiguity aversion in other dimensions, including ambiguity about outcomes.

K-implements f. If f is K-implementable for some K, then f is *implementable*.

K-implementation is the main implementation concept studied in this paper. The IC condition requires the contract to induce beliefs making truthful reporting a best response whenever the agent chooses to participate. The IR condition requires the agent to expect an outcome at least as good as his outside option if he responds truthfully. The final requirement states that the outcome the agent actually receives after reporting θ is $f(\theta)$. Thus, K-implementation is achieved by inducing beliefs making truthful reporting appear to be the safest course of action for all types.

Note that 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, given the worst-case criterion. Arbitrary correspondences b will be called *incentive-compatible* if they satisfy these conditions.

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 $f(\theta) \succsim_\theta \overline{x}_\theta$. This means there is no conflict between the preferences of the agent and the objective of the principal. 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, $\overline{x}_H = 4$, his least-preferred outcome). Suppose the agent is fully rational, so that for any contract he deduces

f (condition 3 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.

3.1 Implementable Functions: Characterization

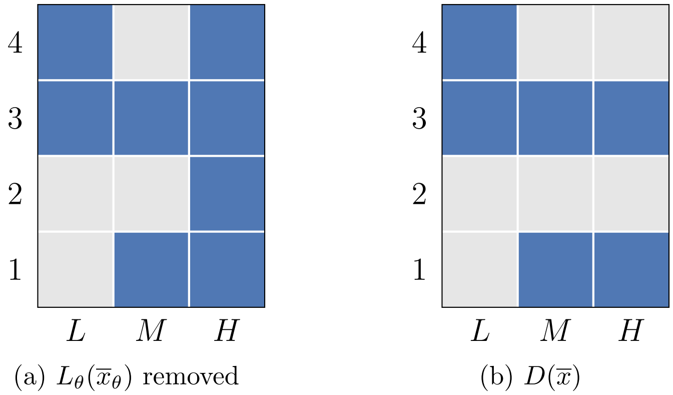
Having defined the agent's cognitive procedure as well as the concept of K-implementation, analysis of the design problem is quite simple. The next definition provides a condition fully characterizing the set of implementable functions. For each $\theta \in \Theta$ and $x \in X$, let $L_\theta(x) := \{y \in X : x \succ_\theta y\}$ 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 $x \in X$, $L_{\theta'}(\overline{x}_{\theta'}) \supseteq L_\theta(x)$ implies $f(\theta) \succsim_\theta x$. Let $D(\overline{x})$ denote the set of all IR-Dominant functions.

IR-Dominance requires type θ to prefer $f(\theta)$ over x whenever types θ and θ' agree that any outcome dominated by x (according to $\succsim_\theta$) is also dominated by $\overline{x}_{\theta'}$, the outside option for θ'. To understand why this is a necessary condition for implementability, suppose b is an incentive-compatible induced belief. Then $b(\theta') \subseteq X \backslash L_{\theta'}(\overline{x}_{\theta'})$ by IR. Therefore, if $L_{\theta'}(\overline{x}_{\theta'}) \supseteq L_\theta(x)$, type θ believes his outcome from misreporting as type θ' is at least as good as outcome x. To satisfy IC, then, he must believe that the worst-case outcome from truthfully reporting type θ is at least as good as x. Since induced beliefs b satisfy $f(\theta) \in b(\theta)$, this forces $f(\theta) \succsim_\theta x$.

Proposition 1.
A function is implementable if and only if it is IR-Dominant.

This result characterizes the set of implementable functions without reference to any details of the agent's bounded rationality (for example, ability K): only



(a) $L_\theta(\overline{x}_\theta)$ removed

(b) $D(\overline{x})$

Figure 2: Constructing $D(\overline{x})$. Panel (a) illustrates a correspondence satisfying the IR condition (outcomes in $L_\theta(\overline{x}_\theta)$ are removed as possible consequences of reporting θ). Notice that type H would prefer to misreport as type M. In (b), outcomes 2 and 4 are removed as possible consequences of report H—the minimal change needed to satisfy IC. The resulting correspondence represents $D(\overline{x})$.

preferences $(\succsim_\theta)_{\theta\in\Theta}$ and outside options $(\overline{x}_\theta)_{\theta\in\Theta}$ are needed to determine if a function f is implementable. Example 3 below provides an illustration.

Proposition 1 is silent regarding the properties of implementing contracts and how they vary with K or f; these issues are examined in the next section. Since the proof of Proposition 1 is closely linked to those results, I defer the proof to section 3.2.1.

Example 3 (continued from Example 2)**.**
Suppose A, Θ, X, preferences, and outside options are as in Example 2. Figure 2 shows how to construct the set of IR-Dominant functions; in particular, $D(\overline{x})$ consists of all functions contained in the largest incentive-compatible belief correspondence. In this case, $f^1 = (4,1,3)$ and $f^3 = (4,3,1)$ are IR-Dominant but $f^2 = (2,2,2)$ and $f^4 = (4,3,2)$ are not. Thus, as suggested in Example 2, f^1 and f^3 are implementable but f^2 and f^4 are not.

3.2 Optimal Contracts

In this section, I show that it is without loss of generality to restrict attention to a stronger implementation concept where contracts K-implement functions for all K up to some bound. Consequently, for any IR-Dominant f, there is an "optimal" contract in the sense that it achieves implementation for the widest possible range of abilities K. Optimal contracts take the following form:

Definition 6 (Complex Contract)**.**
Let $f \in D(\overline{x})$. The *complex contract for* f, denoted $\mathcal{C}_f$, is defined by

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

Each clause of $\mathcal{C}_f$ is formed by taking the set $D(\overline{x})$ and removing a single mechanism $g \in D(\overline{x})$; 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.

As shown in section 3.2.1 below, the set $D(\overline{x})$ qualifies as a belief state. Moreover, the belief state is incentive-compatible. Therefore, $\mathcal{C}_f$ can be interpreted as the result of a simple design principle: maximize the difficulty of performing transitions subject to the constraint that each individual clause makes truthful reporting appear optimal. The main result of this paper establishes that the principal need only consider such contracts:

Proposition 2.
If f is K-implementable, then $\mathcal{C}_f$ K'-implements f for all $K' \leq K$.

Proposition 2 implies $\mathcal{C}_f$ is an optimal contract from the principal's perspective: if some other contract K-implements f, so does $\mathcal{C}_f$. Thus, the principal can do no better than to select $\mathcal{C}_f$ and hope that K is not sufficiently large for the agent to deduce the true mechanism $g_{\mathcal{C}_f} = f$. If K is too large (and f is nontrivial), no other contract would achieve implementation anyway.

Note that $\mathcal{C}_f$ is essentially the unique contract achieving implementation for the full range of admissible K (any other such contract would induce the same beliefs as $\mathcal{C}_f$ for all K, but would contain more clauses and, hence, exhibit some redundancies—see Example 4 below). While it seems natural for contracts to be robust to lower levels of K, this too is a unique feature of $\mathcal{C}_f$: if some other contract K-implements f, then lower-ability agents will arrive at coarser beliefs, but there is no guarantee that those beliefs will be incentive compatible. By employing $\mathcal{C}_f$, the principal nullifies any trade-off between exploitation and robustness: any goal that can be achieved through exploitation of the agent's bounded rationality can be achieved in a way that is robust to heterogeneity in K. In section 3.4, I show that $\mathcal{C}_f$ is also robust to a variety of alternative modeling assumptions (cognitive procedures).

Corollary 1.
There is an integer $\overline{K} \geq 1$ such that, for all nontrivial $f \in D(\overline{x})$, f is K-implementable if and only if $K < \overline{K}$.

This result establishes an upper bound on the agent's cognitive ability beyond which implementation of nontrivial functions is impossible. Below this bound, the complex contract achieves implementation. Note that $\overline{K}$ is independent of f but determined, in part, by the choice of action space $A = \Theta$. In section 3.3.2, I show that $\overline{K}$ grows arbitrarily large (but the set of implementable functions remains the same) as the action space expands.

Example 4 (continued from Examples 2 and 3)**.**
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 $D(\overline{x})$ 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, $\overline{K} = 4$).

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 $B = D(\overline{x})$, whereas in $\mathcal{C}$ the agent must process C_0 in order to reach B.

3.2.1 Proof of Propositions 1 and 2

There are two main steps involved in proving Propositions 1 and 2. First, I show that $D(\overline{x})$ qualifies as a belief state and that it represents the largest incentive-compatible belief correspondence. Then, I show that for all nontrivial $f \in D(\overline{x})$, complex contracts induce beliefs $D(\overline{x})$ whenever some contract K-implements f for some K. From this, the propositions follow quite easily. The arguments presented here are meant to convey the main ideas; a detailed and rigorous proof can be found in the appendix.

Lemma 2.
The set of IR-Dominant functions, $D(\overline{x})$, is a member of $\mathcal{B}$ and satisfies the IC and IR constraints: if $g_{\mathcal{C}} = f$ and $B_{K,\mathcal{C}} = D(\overline{x})$, then $\mathcal{C}$ K-implements f. Moreover, if a contract, $\mathcal{C}$, K-implements a function f, then $B_{K,\mathcal{C}} \subseteq D(\overline{x})$.

Lemma 2 makes three claims. First, it states that $D(\overline{x}) \in \mathcal{B}$; that is, there is a correspondence b^* such that $f(\theta) \in b^*(\theta)$ for all θ if and only if f is IR-Dominant. Second, it states that b^* satisfies the IC and IR constraints. Third, it states that b^* is the largest incentive-compatible belief correspondence: if some other belief b satisfies the constraints, then $b(\theta) \subseteq b^*(\theta)$ for all θ. Thus, if a contract, $\mathcal{C}$, K-implements a function f for some K, then $f \in B_{K,\mathcal{C}} \subseteq D(\overline{x})$, establishing IR-Dominance as a necessary condition for implementability.

The logic of Lemma 2 is illustrated in Figure 2 of Example 3 above. The idea is to construct b^* by first removing outcomes dominated by outside options (thus ensuring IR), and then removing any additional outcomes needed to satisfy IC. That is, for each type θ, set $b^0(\theta) := X \backslash L_\theta(\overline{x}_\theta)$. If b^0 satisfies IC and IR, take $b^* := b^0$. Otherwise, for any type θ that strictly prefers to misreport given b^0, remove only those outcomes $x \in b^0(\theta)$ needed to make type θ (weakly) prefer truthful reporting. Repeating this for all θ yields b^*. It is then a straightforward exercise to verify the claims of Lemma 2.

Lemma 3.
If a contract, $\mathcal{C}$, K-implements f, then $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$.

Lemma 3 says that if $\mathcal{C}$ induces incentive-compatible beliefs, then an agent of ability K forms coarser beliefs under the complex contract $\mathcal{C}_f$. The intuition for this result is as follows. If $\mathcal{C}$ K-implements f, then $B_{K,\mathcal{C}} \subseteq D(\overline{x})$ by Lemma 2. Thus, the state $D(\overline{x})$ is K-reachable under $\mathcal{C}$. Clearly, $D(\overline{x})$ is also K-reachable under $\mathcal{C}_f$ because each $C \in \mathcal{C}_f$ is a subset of $D(\overline{x})$. Since each clause of $\mathcal{C}_f$ eliminates only a single mechanism of $D(\overline{x})$, $\mathcal{C}_f$ maximizes the number of clauses that must be combined in order to transition among subsets of $D(\overline{x})$. Consequently, $\mathcal{C}_f$ yields weakly coarser beliefs for all K.

Lemma 4.
For all K, either $B_{K,\mathcal{C}_f} = D(\overline{x})$ or $B_{K,\mathcal{C}_f} = \{f\}$.

This lemma states a key property of $\mathcal{C}_f$: the agent either deduces f or gets stuck in state $D(\overline{x})$. The reason is quite simple. Since the agent can reach $D(\overline{x})$, the only way to reach a finer belief state is to eliminate some point from the correspondence b^*. For example, to eliminate a point $x \in b^*(\theta)$, the agent must rule out all functions contained in b^* passing through the point (θ, x). There are precisely $\prod_{\theta' \neq \theta} |b^*(\theta')|$ such functions. Since each clause of $\mathcal{C}_f$ eliminates one function, it follows that the agent must have ability

$$K \geq \overline{K} := \min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b^*(\theta')| \tag{2}$$

to eliminate some outcome as a possible consequence of some report. A straightforward induction argument establishes that, in fact, such an agent will successfully pin down f (first he will repeatedly eliminate outcomes from $b^*(\theta)$ until only $f(\theta)$ remains, where θ attains the min in (2); then, an even smaller value of K is required to eliminate outcomes from other reports θ'). Thus, agents of ability $K < \overline{K}$ arrive at beliefs $B_{K,\mathcal{C}_f} = D(\overline{x})$, while those of ability $K \geq \overline{K}$ arrive at $B_{K,\mathcal{C}_f} = \{f\}$.

It follows immediately that every IR-Dominant f is K-implemented by $\mathcal{C}_f$ for some K (namely $K = 1$), proving Proposition 1.[8] For Proposition 2, suppose a contract, $\mathcal{C}$, K-implements a nontrivial f for some K. Then, by Lemma 3, $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$. If $K < \overline{K}$, then $B_{K,\mathcal{C}_f} = D(\overline{x})$ and $\mathcal{C}_f$ K-implements f. If $K \geq \overline{K}$, then $B_{K,\mathcal{C}_f} = \{f\}$. By Lemma 3, this implies $B_{K,\mathcal{C}} = \{f\}$, contradicting the fact that $\mathcal{C}$ K-implements the (nontrivial) function f.

3.3 Comparative Statics

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

3.3.1 Outside Options

Both the set of implementable functions $D(\overline{x})$ and the bound $\overline{K}$ vary with the profile $\overline{x}$ of outside options. In this section only, I will write $\overline{K}(\overline{x})$ to emphasize

[8] If $\overline{K} = 1$, a slightly different argument is needed; see the appendix.

this dependency. The following result is a straightforward consequence of (the proof of) Propositions 1 and 2:

Corollary 2.
If $\overline{x}'_\theta \succsim_\theta \overline{x}_\theta$ for all θ, then $D(\overline{x}') \subseteq D(\overline{x})$ and $\overline{K}(\overline{x}') \leq \overline{K}(\overline{x})$.

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

An interesting special case is when each $\overline{x}_\theta$ 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 $D(\overline{x}) = G$, so that every function is implementable and $\overline{K}(\overline{x}) = |X|^{|\Theta|-1}$. This case still requires the full proof, sketched above, to establish both implementability as well as the optimality of complex contracts.[9]

3.3.2 Larger Action Sets

In this section, I consider action spaces A such that $|A| \geq |\Theta|$. With such action sets, the set of implementable functions is the same but the upper bound $\overline{K}$ increases as $|A|$ increases.

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

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

[9] Note also 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 (eg, large fines), then the agent likely has a more attractive outside option.

It is easy to see that $D^A(\overline{x})$ is the largest incentive-compatible belief correspondence and that $\mathcal{C}_f^A$ K-implements f for all K below a bound $\overline{K}^A(\overline{x})$. In particular, a function is implementable if and only if it is IR-Dominant, and

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

Note that $\overline{K}^A(\overline{x})$ is strictly increasing in the cardinality of A. 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.

3.4 Robustness properties

Propositions 1 and 2 fully characterize the set of implementable functions and establish that $\mathcal{C}_f$ is “optimal”. In this section, I argue that $\mathcal{C}_f$ remains effective under a variety of alternative modeling assumptions. Additional extensions and variations requiring more analysis are presented in section 4.

1. Randomness and Impatience. The cognitive procedure requires the agent to continue processing clauses and performing transitions until further refinement of his beliefs requires the combination of more than K clauses. In principle, arriving at the induced beliefs of Lemma 1 could require many rounds of calculations and transitions. What if the agent is impatient (ie, terminates the procedure prematurely) or processes only some random subset of clauses?

By definition, each clause of $\mathcal{C}_f$ is a subset of $D(\overline{x})$. Therefore, the agent only needs to process one clause of $\mathcal{C}_f$ in order to reach the (incentive compatible) belief state $D(\overline{x})$, and any clause will suffice. Because of this, $\mathcal{C}_f$ is quite robust to the possibilities mentioned above. As long as the agent processes at least one clause (but never $\overline{K}$ or more at a time), $\mathcal{C}_f$ remains effective.

Escaping state $D(\overline{x})$ is difficult not only because it requires the combination of $\overline{K}$ clauses, but because only a small number of sets of $\overline{K}$ clauses actually enable transitions finer belief states. As explained in Example 4, where $\overline{K} = 4$, $\mathcal{C}_f$ contains only three sets of four clauses that enable a transition to a finer belief state. Since $|\mathcal{C}_f| = 7$, this means only three of the $\binom{7}{4} = 35$ subsets of cardinality 4 allow transitions to finer states. Thus, even an agent of ability $K = 4$ will get stuck in $D(\overline{x})$ if he is not patient enough to examine most combinations of four clauses from $\mathcal{C}_f$.

2. Enhanced Deductive Capabilities. The agent's procedure can also be enhanced without severely diminishing the effectiveness of $\mathcal{C}_f$. For example, suppose the agent attempts to perform a "proof by contradiction": he guesses a value of $g_{\mathcal{C}_f}(\theta)$ for some θ, thereby entering a belief state $B \subsetneq D(\overline{x})$, and eliminates this value as a candidate for $g_{\mathcal{C}_f}(\theta)$ only if he can derive a contradiction. More precisely, he rejects his guess if, starting from state B, there is a sequence of K-valid transitions that lead to the empty set.[10] The only effect of introducing this richer procedure is to lower the threshold $\overline{K}$; contract $\mathcal{C}_f$ is still optimal in the sense described above, but now only achieves implementation for a smaller range of abilities.

3. Endogenous *K*. As shown in Lemma 4, behavior under $\mathcal{C}_f$ is "bang-bang": the agent either deduces f (requiring $K \geq \overline{K}$) or gets stuck in state $D(\overline{x})$. Now suppose K were endogenously determined by having the agent assess costs and benefits of different values of K. Let $c(K)$ denote the cost of acquiring ability K (where c is strictly increasing in K and $c(1) = 0$) and consider an agent of type θ. If he chooses $K \geq \overline{K}$, he will deduce f and attain his most-preferred outcome, x^*_θ, in the range of f. If instead he chooses $1 \leq K < \overline{K}$, he will end up holding beliefs $D(\overline{x})$ and reporting truthfully, yielding outcome $f(\theta)$. If $u_\theta(x^*_\theta) - u_\theta(f(\theta)) \geq c(\overline{K})$, then the agent would choose to acquire ability $\overline{K}$. If instead $u_\theta(x^*_\theta) - u_\theta(f(\theta)) < c(\overline{K})$, the agent would choose $K = 1$ (intermediate values of K would all result in outcome $f(\theta)$ but entail cost $c(K) > c(1)$).[11] Thus, the agent's choice of K would

[10] That is, there are states B', B'' that can be reached from B through sequences of K-valid transitions, but $B' \cap B'' \notin \mathcal{B}$.

[11] This assumes the agent correctly assesses the benefit from choosing K (ie, utility stem-

also be "bang-bang", and $\mathcal{C}_f$ would remain optimal from the principal's perspective because it maximizes the level K (and, hence, the cost) required to deduce f. Under $\mathcal{C}_f$, implementation would fail only for those types θ such that $u_\theta(x^*_\theta) - u_\theta(f(\theta)) \geq c(\overline{K})$ and $f(\theta) \neq x^*_\theta$. Intuitively, $u_\theta(x^*_\theta) - u_\theta(f(\theta))$ measures the degree of conflict between type θ and the principal. Whether this conflict is great enough to justify acquiring ability $\overline{K}$ depends on cardinal properties of u_θ and the cost function c. So, while IR-Dominance would still be a necessary condition for implementability (and $\mathcal{C}_f$ the optimal contract), implementation would only be achieved for those types θ and functions f where the conflict between the principal and agent is not too large. As shown in section 3.3.2, however, the principal could make $\overline{K}$ arbitrarily large by inflating the action space (rather than restricting attention to direct mechanisms), restoring IR-Dominance as a sufficient condition for implementability.

4. Ambiguity Attitude. Since beliefs $b_{K,\mathcal{C}}$ typically associate more than one outcome to any given action, the agent's behavior is determined, in part, by his attitude toward ambiguity. The baseline model considers "maxmin" agents who behave as if the worst-possible outcome in $b_{K,\mathcal{C}}(a)$ will result from action a. Perhaps surprisingly, $\mathcal{C}_f$ implements an IR-Dominant function f not only for maxmin agents, but also for agents employing 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 $U^\alpha_\theta(a)$ to action a, where

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

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 K-

ming from choices under induced beliefs for level K) before he actually chooses K. The "circularity" of this approach has obvious conceptual drawbacks, and very similar issues arise in the literature on costly information processing. Nonetheless, correct forecasting of this sort seems to be a natural benchmark. For more on costs and benefits of reasoning, see Alaoui and Penta (2015, 2016).

implements $f \in D(\overline{x})$ for all $K < \overline{K}$ and all $\alpha \in [0,1]$. This is so because beliefs b^* (where $B^{b^*} = D(\overline{x})$) satisfy appropriate IR and IC constraints: IR ($U_\theta^\alpha(\theta) \geq u_\theta(\overline{x}_\theta)$) is satisfied because $\min_{x \in b^*(\theta)} u_\theta(x) \geq u_\theta(\overline{x}_\theta)$, and IC ($U_\theta^\alpha(\theta) \geq U_\theta^\alpha(\theta')$ for all θ') is satisfied because $\operatorname{argmax}_{x \in X} u_\theta(x) \in b^*(\theta)$ and $\min_{x \in b^*(\theta)} u_\theta(x) \geq \min_{x \in b^*(\theta')} u_\theta(x)$ for all θ'. To understand why these inequalities hold, see Figure 2 or the discussion following Lemma 2 above.

4 Extensions and Variations

4.1 Strict Implementation

The definition of K-implementation does not require the agent to strictly prefer truthful reporting. If a contract induces beliefs making the agent indifferent between multiple responses, he may not truthfully report his type unless he suffers a small cost of lying or, more generally, is "white lie averse".[12]

If truth-telling is not sufficiently salient, the principal may wish to design a contract making truthful reporting the unique best response. That is, she may prefer *strict* K-implementation:

> **Definition 7** (Strict K-Implementation).
> Let $K \geq 1$. A contract, $\mathcal{C}$, *strictly K-implements* the function $f : \Theta \to X$ if, for all $\theta, \theta' \in \Theta$ with $\theta \neq \theta'$,
>
> 1. $U_\theta(\theta, K, \mathcal{C}) > U_\theta(\theta', K, \mathcal{C})$ (Strict Incentive Compatibility),
> 2. $U_\theta(\theta, K, \mathcal{C}) \geq u_\theta(\overline{x}_\theta)$ (Individual Rationality), and
> 3. $g_\mathcal{C}(\theta) = f(\theta)$.
>
> A function f is *strictly K-implementable* if there exists a contract that strictly K-implements f. If f is strictly K-implementable for some K, then f is *strictly implementable.*

[12] White lie aversion has recently been applied in other implementation settings. See Matsushima (2008a), Matsushima (2008b), Dutta and Sen (2012), Kartik, Tercieux, and Holden (2014), and Ortner (2015).

This definition replaces the IC condition of K-implementability with Strict Incentive Compatibility: under the induced beliefs, agents who participate in the mechanism strictly prefer truthful reporting.

Definition 8 (Strict IR-Dominance)**.**
A function f *Strictly IR-Dominates* $\overline{x}$ if there is a profile $(x^*_\theta)_{\theta\in\Theta}$ of outcomes such that

1. $f(\theta) \succsim_\theta x^*_\theta \succsim_\theta \overline{x}_\theta$ for all θ, and
2. For all $\theta, \theta' \in \Theta$, $L_\theta(x^*_\theta) \subseteq L_{\theta'}(x^*_{\theta'})$ implies $\theta = \theta'$.

Let $D^*(\overline{x})$ denote the set of Strictly IR-Dominant functions.

This definition says that f is Strictly IR-Dominant if there is a selection of lower-contour sets $L_\theta(x^*_\theta)$ (one for each θ) such that (i) type θ weakly prefers $f(\theta)$ over $\overline{x}_\theta$ and any $y \in L_\theta(x^*_\theta)$, and (ii) $L_\theta(x^*_\theta)$ does not contain $L_{\theta'}(x^*_{\theta'})$ for any $\theta' \neq \theta$. The idea is that any belief correspondence b of the form $b(\theta) = X \backslash L_\theta(x^*_\theta)$ will make truthful reporting the unique optimal response for all types θ.

Indeed, as demonstrated in the appendix, there is a largest such correspondence b^{**}, and the set $D^*(\overline{x})$ coincides with the set of all mechanisms g such that $g(\theta) \in b^{**}(\theta)$ for all θ. Thus, $D^*(\overline{x})$ (when nonempty) is a member of $\mathcal{B}$ and the techniques developed to analyze K-implementation apply for strict K-implementation. In particular, for any $f \in D^*(\overline{x})$, let the *strict complex contract* for f be defined by:

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

This is similar to $\mathcal{C}_f$ in that every clause of $\mathcal{C}^*_f$ reveals $f \in D^*(\overline{x})$ but eliminates only one Strictly IR-Dominant function, thereby maximizing the ability K required to refine beliefs $D^*(\overline{x})$. The following proposition follows from an argument similar to the one developed for K-implementation (see the appendix for details):

Proposition 3.
A function f is strictly implementable if and only if it is Strictly IR-Dominant. Moreover, every such f is strictly K-implementable if and only if $\mathcal{C}_f^*$ strictly K-implements f. Hence, there is an integer $\overline{K}^* \geq 1$ such that, for all nontrivial $f \in D^*(\overline{x})$, f is strictly K-implementable if and only if $K < \overline{K}^*$.

It is easy to see that $D^*(\overline{x}) \subseteq D(\overline{x})$. In other words, Strict IR-Dominance implies IR-Dominance. It follows immediately that $\overline{K}^* \leq \overline{K}$. Thus, the requirement of strict implementation not only shrinks the set of implementable functions, but also the range of abilities K for which implementation can be achieved.

Unlike K-implementation, strict K-implementation rules out some functions f even when agents do not have outside options, and $D^*(\overline{x})$ is empty for some parameter specifications. Nonetheless, the techniques for analyzing strict implementation are nearly identical, and similar results emerge (existence of optimal contracts, robustness properties, and comparative statics).

4.2 Ambiguity Attitude

In this section, I show how to examine the model under alternative assumptions regarding the agent's attitude toward ambiguity. The main takeaway is that ambiguity attitude affects the set of implementable functions, but not the broader lessens derived from the model—a contract with very similar properties to $\mathcal{C}_f$ will be optimal, and therefore similar robustness properties emerge. The impact on the set of implementable functions need not be severe: as demonstrated in section 3.4, $\mathcal{C}_f$ implements an IR-Dominant function f even if the agent employs the Hurwicz α-criterion (for any α).

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

1. 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.

2. The contract
$$\mathcal{C}_f^b = \{B^b \backslash \{g\} : g \in B^b \text{ and } g \neq f\}$$
K-implements f for all $K < \overline{K}^b := \min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b(\theta')|$.

To maximize the range of K for which implementation can be achieved, choose a correspondence b from step 1 that maximizes $\overline{K}^b$. For maxmin agents, this is done by taking $b = b^*$ where $B^{b^*} = D(\overline{x})$. Different ambiguity 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. However, the resulting contract $\mathcal{C}_f^b$ satisfies all of the robustness properties of $\mathcal{C}_f$: it is "optimal" in that it achieves implementation for all K where K-implementation can be achieved by some contract, and it is robust to variations on the cognitive procedure as described in section 3.4. The comparative static results of section 3.3 also carry over.[13]

4.3 Richer 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 9 (Belief system)**.**
A *belief system* is a family $\widehat{\mathcal{B}}$ of subsets of G such that:

B1. Every $B \in \widehat{\mathcal{B}}$ is nonempty.

[13] This procedure could also be used to study implementation for agents who initially hold a prior (probabilistic beliefs) over G, update those beliefs to be consistent with induced belief state B, and then use expected utility to choose a response. For example, an agent with uniform prior on G will end up holding a uniform posterior on B. Clearly, the set of implementable functions would depend on the prior as well as cardinal properties of the utility functions u_θ.

B2. If $B, B' \in \widehat{\mathcal{B}}$ and $B \cap B' \neq \emptyset$, then $B \cap B' \in \widehat{\mathcal{B}}$.

B3. If $B \in \mathcal{B}$, then $B \in \widehat{\mathcal{B}}$.

B1 states that belief states are nonempty sets of mechanisms. B2 says 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 $T = (K, \widehat{\mathcal{B}})$ where $K \geq 1$ is an integer and $\widehat{\mathcal{B}}$ is a belief system. If $T' = (K', \widehat{\mathcal{B}}')$, then $T' \leq T$ means $K' \leq K$ and $\widehat{\mathcal{B}}' \subseteq \widehat{\mathcal{B}}$. 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 $T = (K, \widehat{\mathcal{B}})$, the definitions of K-validity and K-reachability can be adapted from definitions 1 and 2 by replacing $\mathcal{B}$ with $\widehat{\mathcal{B}}$. Call the resulting concepts T*-validity* and T*-reachability*, respectively. Properties B1 and B2 ensure that Lemma 1 holds for arbitrary T (see the appendix). In particular, given a contract $\mathcal{C}$, there is a unique finest T-reachable belief state, denoted $\hat{B}_{T,\mathcal{C}}$. The *effective* belief state, denoted $B_{T,\mathcal{C}}$, is the smallest member of $\mathcal{B}$ containing $\hat{B}_{T,\mathcal{C}}$. Hence, the effective belief state is associated with a correspondence $b_{T,\mathcal{C}}$ given by

$$b_{T,\mathcal{C}}(a) := \{g(a) : g \in \hat{B}_{T,\mathcal{C}}\}$$

The idea of an effective belief state is that if the agent has arrived in state $\hat{B}_{T,\mathcal{C}}$ and if $g \in \hat{B}_{T,\mathcal{C}}$, 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.

Given a contract $\mathcal{C}$, a cognitive type T, and a function $f \in B_{T,\mathcal{C}}$, let

$$\mathcal{C}_{T,f} := \{B_{T,\mathcal{C}} \backslash \{g\} : g \in B_{T,\mathcal{C}},\ g \neq f\}$$

This is similar to the complex contract $\mathcal{C}_f$, but replaces $D(\overline{x})$ 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 4.
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 4 is quite similar to that of Propositions 1 and 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 $\widehat{\mathcal{B}}$ may make the agent highly adept at transitioning away from state $D(\overline{x})$, and therefore $\mathcal{C}_f$ may fail to implement some IR-Dominant f.[14] But if an incentive-compatible state B with $f \in B$ is T-reachable, then the contract $\{B \backslash \{g\} : g \in B,\ g \neq f\}$ will T'-implement f for all $T' \leq T$. Taking $B = B_{T,\mathcal{C}}$ for some $\mathcal{C}$ that T-implements f yields a contract $\mathcal{C}_{T,f}$ with robustness properties similar to $\mathcal{C}_f$.

5 Discussion

5.1 Related Literature

A small but 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 non-standard choice or strategic behavior. For example, Korpela (2012) and de Clippel (2014) study implementation for agents with non-standard choice func-

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

tions, while De Clippel, Saran, and Serrano (2017) and Kneeland (2018) study mechanism design for agents with level-k strategic reasoning (Stahl and Wilson (1994, 1995)).[15] The literature on mechanism design with ambiguity-averse agents (Gilboa and Schmeidler (1989)) also belongs to this category. Bose and Renou (2014) argue 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.[16]

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.[17]

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

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

[16] Di Tillio, Kos, and Messner (2016) show 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).

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

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 acceptance. The agent, rather than forming beliefs and acting on them, adheres to a particular algorithm indicating how his true type interacts with the conditions to generate a response. 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.[18] 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 non-standard 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

[18] 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.

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.

5.2 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: maximize the difficulty of deducing the true mechanism, 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 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 $\widehat{\mathcal{B}}$ 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 $\bigcap_{E_i \in \mathcal{F}} E_i \neq \emptyset$. 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 $\widehat{\mathcal{B}}$ 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 4.3, 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.

A Proofs for Sections 2 and 3

A.1 Preliminaries

This section establishes some basic properties of the cognitive procedure. Since Proposition 4 utilizes the more general model introduced in section 4.3, results are presented for general cognitive types $T = (K, \widehat{\mathcal{B}})$ where $\widehat{\mathcal{B}}$ is a belief system satisfying properties B1–B3 of definition 9. Throughout, $\mathcal{B}$ denotes the baseline belief system where $B \in \mathcal{B}$ if and only if there is a correspondence b such that $B = B^b$.

A transition $B \xrightarrow{\mathcal{C}'} B'$ is *T-valid* (under contract $\mathcal{C}$) if $B, B' \in \widehat{\mathcal{B}}$, $\mathcal{C}' \subseteq \mathcal{C}$ with $|\mathcal{C}'| \leq K$, and

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

Similarly, a state $B \in \widehat{\mathcal{B}}$ is *T-reachable* if there is a sequence

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

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 $T = (\mathcal{B}, K)$.

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

(i) If $B \xrightarrow{\mathcal{C}'} B'$ is T-valid and $B' \subseteq B''$, then $B \xrightarrow{\mathcal{C}'} B''$ is T-valid.

(ii) If B and B' are T-reachable, then $B \cap B' \neq \emptyset$. Hence, $B \cap B' \in \widehat{\mathcal{B}}$.

Proof. For (i), observe that $B \xrightarrow{\mathcal{C}'} B'$ implies $B \cap \left(\bigcap_{C\in\mathcal{C}'} C\right) \subseteq B' \subseteq B''$ and, therefore, $B \xrightarrow{\mathcal{C}'} B''$ is T-valid if $B \xrightarrow{\mathcal{C}'} B'$ is T-valid. For (ii), note that $g_{\mathcal{C}} \in B$ for all T-reachable states B because $\bigcap_{C\in\mathcal{C}} = \{g_{\mathcal{C}}\}$. □

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

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

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 $B, B' \in \widehat{\mathcal{B}}$ are T-reachable, then $B \cap B'$ is T-reachable from B.*

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

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

of T-valid transitions. Observe that

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

Thus,

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

It follows that $B \xrightarrow{\mathcal{C}^1} B \cap \hat{B}^1$ is a T-valid transition. 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 $B \xrightarrow{\mathcal{C}^1} B \cap \hat{B}^1 \xrightarrow{\mathcal{C}^2} B \cap \hat{B}^1 \cap \hat{B}^2 \xrightarrow{\mathcal{C}^3} \ldots \xrightarrow{\mathcal{C}^i}$

$B \cap \hat{B}^1 \cap \ldots \cap \hat{B}^i$ is a sequence of T-valid transitions. Then

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

Thus

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

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

The next result is essentially a restatement of Lemma 1; it establishes that induced beliefs $\hat{B}_{T,\mathcal{C}} \in \hat{B}$ are well-defined for arbitrary types $T = (\widehat{\mathcal{B}}, K)$ and contracts $\mathcal{C}$.

Lemma A.3. *If $\mathcal{C}$ is a contract and $T = (K, \widehat{\mathcal{B}})$, then there is a unique T-reachable state $B^* \in \widehat{\mathcal{B}}$ 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 $\widehat{\mathcal{B}}$ 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, $B' = B^*$. □

Lemma A.4. *If B is T-reachable, then $\hat{B}_{T,\mathcal{C}}$ is the intersection of all $\hat{B} \in \widehat{\mathcal{B}}$ such that $\hat{B}$ is T-reachable from B.*

Proof. Let B be a T-reachable state. As shown in the proof of Lemma A.3, $\hat{B}_{T,\mathcal{C}}$ 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 $\hat{B} = B \cap B'$, as desired. $\square$

A.2 Proof of Propositions 1 and 2

For any $Y \subseteq X$, 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.

Take L^*_θ to be the largest set of the form $L_\theta(Y)$ subject to $Y = L_{\theta'}(\overline{x}_{\theta'})$ ($\theta' \in \Theta'$). That is, there exists θ' such that $L^*_\theta = L_\theta(L_{\theta'}(\overline{x}_{\theta'}))$ and, for all θ'', $L_\theta(L_{\theta''}(\overline{x}_{\theta''})) \subseteq L^*_\theta$.

Let $b^*(\theta) := X \backslash L^*_\theta$. Note that $b^*(\theta) \neq \emptyset$ because no strict lower contour set contains all of X. The following Lemma expands upon Lemma 2 from the main text:

Lemma A.5. *The set of IR-dominant functions, $D(\overline{x})$, satisfies the following:*

(i) $D(\overline{x}) = \{f : \Theta \to X \mid \forall\theta,\ f(\theta) \notin L^*_\theta\}$. *Hence, $D(\overline{x}) \in \mathcal{B}$ and is associated correspondence b^*.*

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

(iii) Every belief state satisfying the IC and IR constraints is a subset of $D(\overline{x})$.

(iv) If $\min_{\theta\in\Theta} \prod_{\theta'\neq\theta} |b^*(\theta')| = 1$, then every $f \in D(\overline{x})$ is trivial.

Proof of (i). Let $B = \{f : \Theta \to X \mid \forall\theta,\ f(\theta) \notin L^*_\theta\}$. I prove that $D(\overline{x}) = B$.

To establish $D(\overline{x}) \subseteq B$, let $f \in D(\overline{x})$. By definition, there is a θ^* such that $L^*_\theta = L_\theta(Y)$ where $Y = L_{\theta^*}(\overline{x}_{\theta^*})$. Since L^*_θ is a strict lower contour of $\succsim_\theta$, there is an $x^* \in X$ such that $L^*_\theta = L_\theta(x^*)$. We have $L_{\theta^*}(\overline{x}_{\theta^*}) \supseteq L_\theta(x^*)$, so that by IR dominance $f(\theta) \succsim_\theta x^*$. 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 $f \in B$ and $L_{\theta'}(\overline{x}_{\theta'}) \supseteq L_\theta(x)$. We need to show that $f(\theta) \succsim_\theta x$. We have $f(\theta) \notin L^*_\theta$, and 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 $L_\theta(x) \subseteq L_\theta(L_{\theta'}(\overline{x}_{\theta'})) \subseteq L^*_\theta$. Thus, $f \in D(\overline{x})$. □

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

Proof of (iii). Suppose B satisfies the IC and IR constraints. Let b denote the associated correspondence, and assume 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 $x \notin L_\theta(\overline{x}_\theta)$ (because $x \in b(\theta) \subseteq X \backslash L_\theta(\overline{x}_\theta)$ by IR). By definition of L^*_θ, there exists θ^* such that $L^*_\theta = L_\theta(L_{\theta^*}(\overline{x}_{\theta^*}))$. We must have $\theta^* \neq \theta$; otherwise, $L^*_\theta = L_\theta(\overline{x}_\theta)$, contradicting the fact that $x \in L^*_\theta \backslash L_\theta(\overline{x}_\theta)$.

Next, observe that if $y \in L_{\theta^*}(\overline{x}_{\theta^*})$, then $y \notin b(\theta^*)$ by the IR constraint for type θ^*. Then $z \notin b(\theta^*)$ for all $z \in L_\theta(L_{\theta^*}(\overline{x}_{\theta^*})) \subseteq L_{\theta^*}(\overline{x}_{\theta^*})$. Thus, under beliefs b, type θ expects (by the worst-case criterion) an outcome strictly better than x from reporting as type θ^*, because $x \in L^*_\theta = L_\theta(L_{\theta^*}(\overline{x}_{\theta^*}))$ and no element of $L_\theta(L_{\theta^*}(\overline{x}_{\theta^*}))$ (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 $g(\theta) \succsim_\theta \overline{x}_\theta$ 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 $g, g' \in D(\overline{x})$, we have $\min_{x \in b^*(\theta^*)} u_{\theta^*}(x) \geq u_{\theta^*}(g(\theta))$ for all $\theta \in \Theta$ and $g \in D(\overline{x})$. Thus, for every $g \in D(\overline{x})$, we have $g(\theta^*) \succsim_{\theta^*} g(\theta)$ for all θ and $g(\theta^*) \succsim_{\theta^*} \overline{x}_{\theta^*}$. Hence, every $g \in D(\overline{x})$ is trivial. □

We now (re)state and prove Lemmas 3 and 4 from the main text.

Lemma 3. *If a contract, $\mathcal{C}$, K-implements f, then $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$.*

Proof. Clearly, state $D(\overline{x}) \in \mathcal{B}$ is K-reachable under $\mathcal{C}_f$ for all K (take $\mathcal{C}' = \{C\}$ for any $C \in \mathcal{C}_f$ to get that $G \xrightarrow{\mathcal{C}'} D(\overline{x})$ is K-valid). By Lemma A.5, $B_{K,\mathcal{C}} \subseteq D(\overline{x})$ and so $D(\overline{x})$ is K-reachable under $\mathcal{C}$ as well. I prove that if $B \xrightarrow{\mathcal{C}'} B'$ is K-valid for some $B, B' \subseteq D(\overline{x})$ and $\mathcal{C}' \subseteq \mathcal{C}_f$, then there is a $\widehat{\mathcal{C}} \subseteq \mathcal{C}$ such that $B \overset{\widehat{\mathcal{C}}}{\to} B'$ is K-valid. This implies that every state that is K-reachable from $D(\overline{x})$ under $\mathcal{C}_f$ is also K-reachable from $D(\overline{x})$ under $\mathcal{C}$. Then $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$ by Lemma A.4.

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

$$B \cap \left(\bigcap_{C \in \mathcal{C}'} C \right) \subseteq B' \qquad (4)$$

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 $\widehat{\mathcal{C}} = \{C^i : i = 1, \ldots, n\}$ and observe that

$B \cap C^i \subseteq D(\overline{x}) \backslash \{g_i\}$. Then

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

Combined with (4), it follows that

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

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

Lemma 4. *For all K, either $B_{K,\mathcal{C}_f} = D(\overline{x})$ or $B_{K,\mathcal{C}_f} = \{f\}$.*

Proof. Let $K \geq 1$. As argued in the proof of Lemma 3, $D(\overline{x}) \in \mathcal{B}$ is K-reachable under $\mathcal{C}_f$. Thus, $B_{K,\mathcal{C}_f} \subseteq D(\overline{x})$. I prove that if some $B \in \mathcal{B}$ such that $B \subsetneq D(\overline{x})$ is K-reachable, then $B_{K,\mathcal{C}_f} = \{f\}$.

Let $b^* := b^{D(\overline{x})}$ denote the correspondence associated with $D(\overline{x})$. For each $\theta \in \Theta$, let $|b^*(\theta)|$ denote the cardinality of $b^*(\theta)$ and note that $|D(\overline{x})| = \prod_{\theta \in \Theta} |b^*(\theta)|$.

If some $B \subsetneq D(\overline{x})$ is K-reachable, then there exist $\mathcal{C}^1, \ldots, \mathcal{C}^n \subseteq \mathcal{C}_f$ and $B^1, \ldots, B^n \in \mathcal{B}$ such that

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

is a sequence of K-valid transitions. We may assume $B^i \subseteq D(\overline{x})$ for all $i \geq 1$ since $C \subseteq D(\overline{x})$ for all $C \in \mathcal{C}_f$. Let i^* be the smallest i such that $B^i \subsetneq D(\overline{x})$ and let $B' = B^{i^*}$.

Letting $\mathcal{C}' = \mathcal{C}^{i^*}$, it follows from our choice of i^* that $D(\overline{x}) \xrightarrow{\mathcal{C}'} B'$ is K-valid. Moreover, since $B' \subsetneq D(\overline{x})$, 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 $\mathcal{C}' = \{D(\overline{x}) \backslash \{g'\} : g' \in E\}$ for some E containing every $g' \in D(\overline{x})$ such that $g'(\theta) = x$. Thus, since $|\mathcal{C}'| \leq K$,

$$|\{g \in D(\overline{x}) : g(\theta) = x\}| \leq K \tag{5}$$

Clearly, (5) holds for every choice of $x \in b^*(\theta)$ such that $x \neq g_{\mathcal{C}_f}(\theta)$ because $|\{g \in D(\overline{x}) : g(\theta) = x\}| = \prod_{\theta' \neq \theta} |b^*(\theta')|$, which does not depend on x.

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

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

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

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

is a sequence of K-valid transitions where, for every $i = 1, \ldots, m$, $\hat{B}^i$ satisfies $b^{\hat{B}^i}(\theta) = b^*(\theta) \backslash \{x_1, \ldots, x_i\}$. The transitions are K-valid because $|\mathcal{C}^{(\theta, x_i)}| = |\{g \in D(\overline{x}) : g(\theta) = x\}|$, which does not exceed K by (5).

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 $B_{-\Theta'} := \{g \in D(\overline{x}) : \forall \theta' \in \Theta',\ g(\theta') = g_{\mathcal{C}_f}(\theta')\}$. 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 $\theta' \in \Theta \backslash \Theta'$.

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

$$\begin{aligned}
|\{g \in B_{-\Theta'} : g(\theta') = x'\}| &= \prod_{\hat{\theta} \in \Theta \backslash (\Theta' \cup \theta')} |b^*(\hat{\theta})| \\
&\leq \prod_{\hat{\theta} \in \Theta \backslash \theta} |b^*(\hat{\theta})| \qquad \text{since } \theta \in \Theta' \\
&\leq K \qquad \text{by (5)}
\end{aligned}$$

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

$$\widehat{\mathcal{C}}^{(\theta',x')} := \{D(\overline{x}) \backslash \{g\} : g \in B_{-\Theta'} \text{ and } g(\theta') = x'\} \subseteq \mathcal{C}_f$$

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

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

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. □

A.2.1 Proof of Proposition 1

Suppose $f : \Theta \to X$ is implementable. If f is nontrivial, then f must be K-implemented (for some K) by a contract $\mathcal{C}$ such that $B_{K,\mathcal{C}} \subseteq D(\overline{x})$. This is so because $B_{K,\mathcal{C}}$ must satisfy the IR and IC constraints, and by Lemma A.5 such beliefs are necessarily a subset of $D(\overline{x})$. Hence, f is IR-Dominant (clearly, trivial functions are IR-Dominant as well).

Conversely, let $f \in D(\overline{x})$. If f is trivial, the contract $\mathcal{C} = \{f\}$ will suffice. Otherwise, consider the complex contract $\mathcal{C}_f$. Let $K = 1$. By Lemma 4, either $B_{K,\mathcal{C}_f} = D(\overline{x})$ or $B_{K,\mathcal{C}_f} = \{f\}$. If $B_{K,\mathcal{C}} = D(\overline{x})$, then $\mathcal{C}_f$ K-implements f. Otherwise, $B_{K',\mathcal{C}_f} = \{f\}$ for all $K' \geq 1$. In particular, an agent of ability $K = 1$ pins down the true mechanism $g_{\mathcal{C}_f} = f$. This can only happen if $\min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b^*(\theta')| = 1$ (because this condition must be satisfied for an agent of ability $K = 1$ to be able to reach finer beliefs than $D(\overline{x})$, thus triggering

Lemma 4). Thus, by part (iv) of Lemma A.5, f is trivial. Hence, in all cases, f is implementable.

A.2.2 Proof of Proposition 2 and Corollaries 1 and 2

Let $\overline{K} := \min_{\theta\in\Theta} \prod_{\theta'\neq\theta} |b^*(\theta')|$ and let $f \in D(\overline{x})$. Observe that ability $K' \geq \overline{K}$ is required for the agent to be able to reach a belief state $B \subsetneq D(\overline{x})$ in contract $\mathcal{C}_f$. Let $\mathcal{C}$ be a contract that K-implements f.

If f is trivial, there is nothing to prove since, by Lemma 4, either $B_{K,\mathcal{C}} = D(\overline{x})$ or $B_{K,\mathcal{C}} = \{f\}$ and both beliefs are incentive-compatible.

If f is nontrivial, apply Lemma 3 to get $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$. If $K < \overline{K}$, then $B_{K,\mathcal{C}_f} = D(\overline{x})$ because only an agent of ability $K' \geq \overline{K}$ can transition from beliefs $D(\overline{x})$ to a proper subset of $D(\overline{x})$ (and, by Lemma 4, $\mathcal{C}_f$ can only induce beliefs $D(\overline{x})$ or $\{f\}$). By Lemma A.5, beliefs $D(\overline{x})$ satisfy the IC and IR constraints, and therefore $\mathcal{C}_f$ K'-implements for all $K' < \overline{K}$, including K.

If $K > \overline{K}$, then $B_{K,\mathcal{C}_f} = \{f\}$; thus, $B_{K,\mathcal{C}} = \{f\}$ by Lemma 3. This contradicts the fact that $\mathcal{C}$ K-implements the (nontrivial) function f, proving Proposition 2. Corollary 1 follows immediately using the above formula for $\overline{K}$.

For corollary 2, observe that if $\overline{x}'_\theta \succsim_\theta \overline{x}_\theta$ for all θ, then $L_\theta(\overline{x}'_\theta) \supseteq L_\theta(\overline{x}_\theta)$ for all θ and, hence, L^*_θ is larger under $\overline{x}'$ than $\overline{x}$ (for all θ). It follows that $D(\overline{x}') \subseteq D(\overline{x})$. Thus, if b' and b are the belief correspondences associated with $D(\overline{x}')$ and $D(\overline{x})$, respectively, then $b'(\theta) \subseteq b(\theta)$. Therefore,

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

B Proofs for Section 4

B.1 Proof of Proposition 3

There is a simple algorithm for determining the set D^* of strictly implementable functions. This is accomplished by constructing the largest correspondence b^{**} satisfying IR and Strict IC, then taking D^* to be the set of all mechanisms contained in b^{**}. The algorithm for b^{**} proceeds as follows:

1. For each θ, remove the sets L^*_θ as possible outcomes from reporting θ so that the resulting correspondence b^0 satisfies $b^0(\theta) = X \backslash L^*_\theta$. If b^0 induces strict preferences for truthful reporting, take $b^{**} = b^0$. If not, proceed to step 2.
2. For each θ such that truthful reporting is not the unique optimal response under beliefs b^i, remove the worst remaining outcome at coordinate θ according to the preferences $\succsim_\theta$. Let b^{i+1} denote the resulting correspondence.
3. If b^{i+1} induces strictly optimal truth telling for all types, take $b^{**} = b^{i+1}$. If not, repeat step 2 with $i + 1$ in place of i.

It is easy to see that this algorithm terminates, but it need not be the case that b^{**} is nonempty-valued. Let $D = \{f : \Theta \to X \mid \forall \theta \in \Theta,\ f(\theta) \in b^{**}(\theta)\}$.

Lemma B.1. *If $f : \Theta \to X$ satisfies $f(\theta) \in b^{**}(\theta)$ for all θ, then f is strictly implementable.*

Proof. By construct, the correspondence b^{**} satisfies IR and Strict IC. Moreover, by a similar argument to that of the previous section, the contract

$$\mathcal{C} = \{D \backslash \{g\} \mid g \in D \text{ and } g \neq f\}$$

either induces beliefs $b_{K,\mathcal{C}} = D$ or $b_{K,\mathcal{C}} = \{f\}$. Clearly, $\mathcal{C}$ strictly K-implements f if $b_{K,\mathcal{C}} = D$ for some K. If $b_{K,\mathcal{C}} = \{f\}$ for all K, then (by a similar argument to part (iv) of Lemma A.5), f is trivially strictly implementable: for all θ, $f(\theta) \succ_\theta f(\theta')$ and $f(\theta) \succsim_\theta \overline{x}_\theta$. □

Lemma B.2. *Any correspondence b satisfying IR and Strict IC is contained in b^{**}; that is, $b(\theta) \subseteq b^{**}(\theta)$ for all θ.*

Proof. Suppose toward a contradiction that b is not a sub-correspondence of b^{**}. Since b satisfies the IR and (regular) IC constraints, we have $b(\theta') \subseteq b^0(\theta')$ for all θ' (Lemma A.5). Hence, there is a smallest $i \geq 0$ such that $b(\theta') \subseteq b^i(\theta')$ for all θ' but $b(\theta) \not\subseteq b^{i+1}(\theta)$ for some θ. Then there is an outcome x such that

$x \in b(\theta) \cap b^i(\theta)$ but $x \notin b^{i+1}(\theta)$. By definition of Step 2 of the algorithm, x minimizes u_θ on the set $b^i(\theta)$, and x gets removed from $b^i(\theta)$ (when forming b^{i+1}) because there is some $\theta' \neq \theta$ and $x' \in b^i(\theta')$ such that (i) x' minimizes u_θ on the set $b^i(\theta')$, and (ii) $x' \succsim_\theta x$. In other words, for type θ, beliefs b^i make response θ' at least as attractive as response θ. Note that since $b(\theta) \subseteq b^i(\theta)$, x also minimizes u_θ on the set $b(\theta)$. There are two cases:

1. If $x' \in b(\theta')$, then x' minimizes u_θ on $b(\theta')$ because $b(\theta') \subseteq b^i(\theta')$ and x minimizes u_θ on $b^i(\theta')$. Thus, type θ weakly prefers reporting θ' over θ under beliefs b, contradicting the fact that b satisfies Strict IC.
2. If $x' \notin b(\theta')$, then type θ prefers any minimizer of u_θ on $b(\theta')$ over x' (the minimizer of u_θ on $b^i(\theta') \supseteq b(\theta')$). Thus, type θ prefers reporting θ over θ' under beliefs b, contradicting the fact that b satisfies Strict IC.

Thus, b is a sub-correspondence of b^{**}. □

It follows immediately from Lemmas B.1 and B.2 that D^*, the set of strictly implementable functions, satisfies

$$D^* = \{f : \Theta \to X \mid \forall \theta \in \Theta,\ f(\theta) \in b^{**}(\theta)\} \tag{6}$$

Next observe that if a function f is Strictly IR-Dominant, then there is a correspondence b containing f that satisfies IR and Strict IC. Specifically, take $b(\theta) := X \backslash L_\theta(\overline{x}^*_\theta)$ where x^* is the profile of outcomes asserted by Strict IR-Dominance. Hence, by Lemma B.1, every $f \in D^*(\overline{x})$ is strictly implementable.

Conversely, by (6), every $f \in D^*$ is a member of $D^*(\overline{x})$ because the algorithm yields a b^{**} of the form $b^{**}(\theta) = X \backslash L_\theta$ where L_θ is a strict lower contour of $\succsim_\theta$. Hence, the desired profile x^* can be found by letting x^*_θ be any minimizer of u_θ over the set $b^{**}(\theta)$.

Thus, $D^*(\overline{x}) = D^*$. The remainder of the argument is analogous to that of Proposition 2 and its corollaries.

B.2 Proof of Proposition 4

Suppose $T = (K, \widehat{\mathcal{B}})$ and that $\mathcal{C}$ T-implements a function f. Let $\hat{B}$ denote the effective belief for T given $\mathcal{C}$, and let $\mathcal{C}_{T,f} := \{\hat{B}\backslash\{g\} : g \in \hat{B}\backslash\{f\}\}$. Let $T' = (K', \widehat{\mathcal{B}}') \leq T$. We must have $f \in \hat{B} \subseteq D(\overline{x})$ since $\hat{B}$ is incentive-compatible.

Lemma B.3. *If there exist $B, B' \in \widehat{\mathcal{B}}'$ such that $B, B' \subseteq \hat{B}$ and $\mathcal{C}' \subseteq \mathcal{C}_{T,f}$ such that the transition $B \xrightarrow{\mathcal{C}'} B'$ is T'-valid, then there exists $\mathcal{C}'' \subseteq \mathcal{C}$ such that $B \xrightarrow{\mathcal{C}''} B'$ is T-valid.*

Proof. Suppose $B, B' \in \widehat{\mathcal{B}}$ (hence, $B, B' \in \widehat{\mathcal{B}}$) and that $B \xrightarrow{\mathcal{C}'} B'$ 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 $\hat{B}\backslash\{g\}$, where $g \in \hat{B}\backslash\{f\}$. Thus, $\bigcap_{C' \in \mathcal{C}'} C' = \hat{B}\backslash\{g_1, \dots, g_{\hat{K}}\}$ where $\hat{K} \leq K'$ and $f \neq g_i \in \hat{B}$ for all $i = 1, \dots, \hat{K}$. 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 $\hat{B} \cap C_i \neq \emptyset$ because $f \in \hat{B} \cap C_i$. Finally, let $\mathcal{C}'' = \{C_1, \dots, C_{\hat{K}}\}$ and observe that $|\mathcal{C}''| \leq \hat{K}$ (with strict inequality if $C_i = C_j$ for some $i \neq j$). Then $B \cap \bigcap_{C'' \in \mathcal{C}''} C'' \subseteq B \cap \bigcap_{C' \in \mathcal{C}'} C' \subseteq B'$. □

This lemma implies that if a state B is T'-reachable from $\hat{B}$ ($T' \leq T$), then it also T-reachable from $\hat{B}$. Since $\hat{B}$ is T'-reachable for all T' under both $\mathcal{C}$ and $\mathcal{C}_{T,f}$, this implies $\mathcal{C}_{T,f}$ induces coarser beliefs than $\mathcal{C}$ does (Lemma A.4) whenever $T' \leq T$.

Now observe that since $\hat{B}$ is T'-reachable under $\mathcal{C}_{T,f}$, the effective belief for T' under $\mathcal{C}_{T,f}$ must be $\hat{B}$ (if it were finer, then $\mathcal{C}$ would have a finer effective belief for T). Thus, $\mathcal{C}_{T,f}$ T'-implements f for all $T' \leq T$.

References

Alaoui, L. and A. Penta (2015). Endogenous depth of reasoning. *The Review of Economic Studies 83*(4), 1297–1333.

Alaoui, L. and A. Penta (2016). Cost-benefit analysis in reasoning. *Working paper*.

Battigalli, P. and G. Maggi (2002). Rigidity, discretion, and the costs of writing contracts. *American Economic Review 92*(4), 798–817.

Bodoh-Creed, A. L. (2012). Ambiguous beliefs and mechanism design. *Games and Economic Behavior 75*(2), 518–537.

Bose, S. and A. Daripa (2009). A dynamic mechanism and surplus extraction under ambiguity. *Journal of Economic theory 144*(5), 2084–2114.

Bose, S., E. Ozdenoren, and A. Pape (2006). Optimal auctions with ambiguity. *Theoretical Economics 1*(4), 411–438.

Bose, S. and L. Renou (2014). Mechanism design with ambiguous communication devices. *Econometrica 82*(5), 1853–1872.

de Clippel, G. (2014). Behavioral implementation. *The American Economic Review 104*(10), 2975–3002.

De Clippel, G., R. Saran, and R. Serrano (2017). Level-k mechanism design. *The Review of Economic Studies (Forthcoming)*.

Di Tillio, A., N. Kos, and M. Messner (2016). The design of ambiguous mechanisms. *The Review of Economic Studies*.

Dutta, B. and A. Sen (2012). Nash implementation with partially honest individuals. *Games and Economic Behavior 74*(1), 154–169.

Eliaz, K. (2002). Fault tolerant implementation. *The Review of Economic Studies 69*(3), 589–610.

Eliaz, K. and P. Ortoleva (2015). Multidimensional ellsberg. *Management Science 62*(8), 2179–2197.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 643–669.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics 18*(2), 141–153.

Glazer, J. and A. Rubinstein (2012). A model of persuasion with boundedly rational agents. *Journal of Political Economy 120*(6), 1057–1082.

Glazer, J. and A. Rubinstein (2014). Complex questionnaires. *Econometrica 82*(4), 1529–1541.

Hurwicz, L. (1951). Some specification problems and applications to econometric models. *Econometrica 19*(3), 343–44.

Kartik, N., O. Tercieux, and R. Holden (2014). Simple mechanisms and preferences for honesty. *Games and Economic Behavior 83*, 284–290.

Kneeland, T. (2018). Mechanism design with level-k types: Theory and an application to bilateral trade. *Working paper*.

Korpela, V. (2012). Implementation without rationality assumptions. *Theory and decision 72*(2), 189–203.

Koszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature 52*(4), 1075–1118.

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review 107*(11), 3257–87.

Lipman, B. L. (1999). Decision theory without logical omniscience: Toward an axiomatic framework for bounded rationality. *The Review of Economic Studies 66*(2), 339–361.

Matsushima, H. (2008a). Behavioral aspects of implementation theory. *Economics Letters 100*(1), 161–164.

Matsushima, H. (2008b). Role of honesty in full implementation. *Journal of Economic Theory 139*(1), 353–359.

Ortner, J. (2015). Direct implementation with minimally honest individuals. *Games and Economic Behavior*.

Salant, Y. and A. Rubinstein (2008). (A, f): Choice with frames. *The Review of Economic Studies 75*(4), 1287–1296.

Salant, Y. and R. Siegel (2018). Contracts with framing. *American Economic Journal: Microeconomics 10*(3), 315–46.

Stahl, D. O. and P. W. Wilson (1994). Experimental evidence on players' models of other players. *Journal of economic behavior & organization 25*(3), 309–327.

Stahl, D. O. and P. W. Wilson (1995). On players models of other players: Theory and experimental evidence. *Games and Economic Behavior 10*(1), 218–254.

Wolitzky, A. (2016). Mechanism design with maxmin agents: Theory and an application to bilateral trade. *Theoretical Economics 11*(3), 971–1004.