A Model of Complex Contracts

Alexander M. Jakobsen*

February 2018

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 mapping actions to outcomes. In this setting, only a small class of trivial functions are implementable if the agent is fully rational. I introduce a model of bounded rationality where the agent has a limited ability to combine different pieces of information. Specifically, the agent transitions between belief states by combining his current beliefs with up to K pieces of information at a time. By expressing the mechanism as a *complex contract*—a collection of clauses, each providing some information about the mechanism—the principal manipulates the agent into believing that the mechanism is ambiguous. I assume the agent is averse to this perceived ambiguity and characterize the set of implementable functions. The main result shows that, without loss of generality, the principal selects a robust, maximally complex contract achieving implementation for all abilities below a bound $\overline{K}$. The model of bounded rationality introduced here extends naturally to other environments and provides an intuitive notion of what it means for a problem to be difficult to solve.

*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 patience, guidance, and support at all stages of this project. Thanks also to Dilip Abreu, Roland Benabou, Sylvain Chassang, Stephen Morris, Doron Ravid, Ariel Rubinstein, Kai Steverson, and Benjamin Young for helpful conversations and feedback.

1 Introduction

In this paper, I examine how a sophisticated designer may benefit from introducing unnecessary complexity into the description of a mechanism. I consider a principal-agent setting where the principal seeks to implement a function mapping agent types to outcomes. The principal commits to a mechanism by announcing a *contract*: a set of clauses that, all combined, define a function (a *mechanism*) assigning outcomes to actions. Each clause provides information about the mechanism, and the agent must process and combine these clauses in order to form beliefs about the mechanism. The main results show that with boundedly rational agents, the principal benefits from introducing excessive complexity into the contract: a well-crafted complex contract manipulates agents into believing that truthful reporting is optimal, thereby expanding the set of implementable functions.

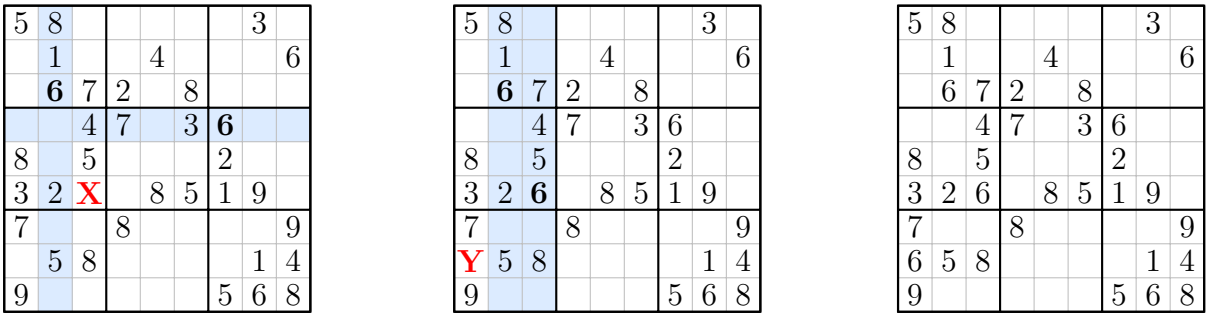
The presence of extreme complexity in contracts and institutions is well-known. Tax codes in particular, and legal systems in general, are notoriously complex. Insurance plans, employment contracts, and end-user license agreements are also familiar examples of the many scenarios in which real people confront highly complex rules and systems.

One explanation for such complexity is that the designer wishes to implement a very nuanced and detailed objective function, so that complexity arises out of the need to articulate, unambiguously, many different cases and contingencies. In my model, the principal has sufficient expressive power that this is not the case: she has the ability to clearly communicate the mechanism in a way that all agents understand, but chooses not to do so in order to implement a wider range of functions.

The fact that people struggle with complexity presents an interesting challenge for economic theorists. Standard approaches assume that economic agents are perfect problem solvers who—regardless of the complexity of a decision problem—effortlessly deduce the optimal course of action.[1] Clearly, real people do not exhibit this characteristic. Is there a systematic way in which they differ from standard, fully-rational agents that is amenable to economic analysis? A primary goal of this paper is to introduce a plausible and intuitive model of bounded cognition capturing some relevant aspects of what it means for a problem to be difficult for a human to solve. The driving force behind the model is that it is difficult to combine many pieces of information (rules) at once.

To illustrate the cognitive procedure, consider the game Sudoku. In this game, a player is presented with a 9×9 square. 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

[1] In particular, agents are implicitly assumed to be *logically omniscient*: if a collection of facts is known to an agent, then so are all of its logical implications.



(a) Player deduces X = 6 (b) . . .then Y = 6 (c) New configuration

Figure 1: A possible sequence of deductions in Sudoku

are that each digit $d \in D$ must appear exactly once in (i) each row; (ii) each column; and (iii) each of the nine main 3×3 subsquares. The puzzles are designed so that there is a unique solution given the initial partially-filled square.

For a standard rational agent, there is no distinction between a partially-filled square—together with knowledge of the rules of the game—and the unique fully-resolved puzzle. To him, a partially-filled square plus the rules of the game simply form a compact way of expressing an entry for each cell. Not so for most (real) people, who understand both the rules of the game as well as the initial configuration but may find themselves unable to solve the puzzle.

How might an individual go about solving a Sudoku puzzle? Consider Figure 1. Suppose the player notices the entry 6 in positions (3,2) and (4,7). From this, he knows that 6 cannot appear again in column 2 or row 4 (Figure 1a). Combined with the fact that 6 must appear in each main 3×3 subsquare, he deduces that X (position (6,3)) must be 6. Hence, he updates the configuration (Figure 1b) to reflect this. Looking at his new configuration, he realizes that 6 cannot appear again in columns 2 or 3. Combined with the fact that 6 must appear somewhere in the bottom left 3×3 subsquare, 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".

What, then, distinguishes a hard puzzle from a simple one? I propose the following. In a simple puzzle, the player is able to chip away at the problem: he can gradually fill in the cells, one at a time, without ever having to combine very many rules at once in order to deduce the entry for another cell. In the example above, the player only had to combine three rules (together with his initial knowledge) to deduce X = 6, and three again to deduce Y = 6 once he updated the configuration. In a hard puzzle, however, the player inevitably reaches a configuration where the only way to fill in another cell is to combine many rules at once; that is, he must perform a large "leap of logic" in order to make progress. If the

player cannot perform the required chain of reasoning, he will remain stuck at the current configuration. If he somehow manages to fill in another square, the remainder of the puzzle may or may not be simple; either way, he must be sophisticated enough to overcome this hurdle in order to proceed.

The model formalizes this idea by stipulating that the agent has a *deductive ability* $K \geq 1$ (an integer) and uses his ability, together with the clauses announced by the principal, to transition between *belief states*. A belief state represents a level of knowledge, and an agent in state B may combine up to K clauses with knowledge B to transition to some other B'. In the Sudoku example, a belief state is a partially-filled table; in the implementation model, a belief state is a set of mechanisms. A clause C signals that the mechanism belongs to C, while a belief state B indicates that it belongs to B. By intersecting up to K clauses with B, the agent may be able to deduce that the mechanism belongs to some other set B'. He continues in this fashion until he is unable to further refine his beliefs.

Notice that in the Sudoku example, the player typically does not retain all new information he has derived when transitioning to new states. For example, when updating his configuration to reflect X = 6, he "forgets" that 6 has been eliminated from column 2 and row 4. This is a crucial element of his bounded rationality; if he always retains all new information, his limited ability to combine rules has no effect and he will be able to solve any puzzle. Belief states capture this forgetfulness; they represent what the agent is able to recall and reason about, or statements that he attempts to prove when performing calculations. He temporarily learns more when combining rules, but only retains information that can be encoded in a belief state.

Accordingly, the agent in the implementation model works with a particular collection of belief states. With finite sets A and X representing action and outcome spaces, respectively, the agent forms a *belief correspondence* $b : A \rightrightarrows X$. I assume there is a one-to-one mapping between belief correspondences and belief states. Thus, for each state B, there is a correspondence b such that B consists of all mechanisms $g : A \to X$ contained in b; that is, $B = \{g \mid \forall a \in A,\ g(a) \in b(a)\}$. The interpretation of a belief correspondence b is that the agent has narrowed down the set of possible outcomes from action a to $b(a)$. In other words, he only reasons about the set of possible outcomes for each action, and only remembers information gleaned from a collection of clauses if it allows him to eliminate some outcome as a possible consequence of some action.

Given the agent's cognitive procedure, the principal sets out to design a contract $\mathcal{C}$, which is a collection of clauses (each a set of mechanisms) that pin down a mechanism: there is a mechanism $g_{\mathcal{C}} : A \to X$ such that $\bigcap_{C \in \mathcal{C}} C = \{g_{\mathcal{C}}\}$. Presented with $\mathcal{C}$, the agent forms a belief correspondence $b_{K,\mathcal{C}}$ and takes the best possible action given these beliefs (if none are

sufficiently attractive, he opts to retain an outside option). If he takes action a, he receives outcome $g_{\mathcal{C}}(a)$, the outcome actually prescribed by $\mathcal{C}$. The principal aims to design $\mathcal{C}$ in such a way that an agent of type $\theta \in \Theta$ obtains outcome $f(\theta)$.

In this setting, only a small class of trivial functions f are implementable for fully rational agents: such agents always deduce the true mechanism and pick their favorite outcome $g_{\mathcal{C}}(a)$ which, except in special cases, does not coincide with $f(\theta)$. The main result shows that if agents are boundedly rational as outline above, then a much larger set of functions is implementable. Moreover, the principal can do no better than to offer a *complex contract* $\mathcal{C}_f$ that achieves robust implementation: it implements an admissible objective function f as long as the agent's ability does not exceed some threshold $\overline{K}$. Beyond this threshold, no contract can implement a nontrivial function. Hence, the principal does not require—and cannot benefit from—any knowledge of the agent's ability K: she simply offers $\mathcal{C}_f$ since it achieves implementation for the full range of abilities below $\overline{K}$.

Since the agent forms a belief correspondence, he perceives the mechanism to be ambiguous: from his perspective, multiple outcomes are possible consequences of a given action. Therefore, an additional assumption is needed regarding his attitude toward this perceived ambiguity. In the main model, I assume he is averse to perceived ambiguity—he behaves as if the worst-possible outcome will attain. I discuss this assumption in section 2.3 and show how to analyze the model under alternative assumptions in section 4.3. Roughly speaking, many insights of the model hold under alternative assumptions because the process of forming beliefs is independent of how the agent evaluates ambiguity. Ambiguity attitude affects the set of implementable functions, but not the finding that the set of implementable functions expands under bounded rationality, nor the result that the principal optimally selects a complex contract achieving robust implementation.

1.1 Related Literature

This paper adds to the growing literature on behavioral mechanism design; see Koszegi (2014) for a survey. Some recent contributions that are more closely related to this paper are Salant and Siegel (2013), who study a monopolist seeking to increase profits by exploiting framing effects, and Korpela (2012) and de Clippel (2014), who consider an implementation setting where agents have nonstandard choice functions. De Clippel, Saran, and Serrano (2014) study mechanism design for agents with level-k strategic reasoning (Stahl and Wilson (1994, 1995)), and Eliaz (2002) considers an implementation setting where some players are error-prone and may fail to behave optimally.

The closest work, however, is a pair of papers by Glazer and Rubinstein (2012, 2014)

(henceforth GR12/14), who study persuasion with boundedly rational agents. In both models, all agents (regardless of type) wish to have a request granted by a principal while the principal only wants to grant the request for a particular subset $\mathcal{A}$ of types. Both papers employ a particular syntactical framework for modeling bounded rationality, but differ in the manner in which agents are bounded as well as the implementation objective faced by the principal. In GR12, the principal specifies a set of conditions using the syntactical framework. These conditions define $\mathcal{A}$ and the agent, instead of 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 a limited ability to detect patterns in the set of acceptable responses. The same syntactical structure is needed to define the patterns that agents detect. Agents are classified as either truthful or manipulative; truthful agents answer all questions truthfully, while manipulative agents attempt to use their pattern recognition ability to choose an acceptable response. Given this classification, the principal solves a constrained implementation problem where all truthful, acceptable types must be accepted while minimizing the probability that manipulators are accepted. They show that this probability depends only on the cardinality of $\mathcal{A}$ and that it decreases very quickly as the set $\mathcal{A}$ grows.

The model and results of this paper differ from GR12 and GR14 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 (although, for ease of exposition, I focus on separating contracts inducing different types to give different responses). 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 characterize not only the set of implementable implementable functions, but also establish that bounds on agent ability are necessary.

This paper is also related to the literature on mechanism design with ambiguity-averse agents (Gilboa and Schmeidler (1989)). 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 that engineers 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.[2]

Lipman (1999) develops an axiomatic model accommodating logical non-omniscience and argues that such bounded rationality stems from a sensitivity to how information is framed. The key difference between Lipman's framework and the deduction model presented here is that while Lipman does not assume a specific reasoning procedure, his framework rules out the possibility that the combination of different pieces of information may be more informative than the individual pieces. His framework also excludes models of resource-bounded reasoning, of which mine is a particular case.

Finally, this paper is also related to the literature on choice with frames. Salant and Rubinstein (2008) (SR) study a model of choice in which a decision maker is presented with a pair (C, f), where C is a set of outcomes and f is a frame for C (for example, an ordering of the elements of C). My framework can be modified so that contracts just describe sets of outcomes and thereby act as frames. The essential difference between my framework and that of SR is that in my model, the decision maker only sees the frame (contract), and may miscalculate the set from which he is choosing (Result 4 below uses this idea).

2 Model

2.1 Outcomes, Types, Contracts

Let Θ denote a finite set of *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.

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 provides information about a mechanism by describing one of its properties. For example, $C = \{g \in G : g(a_3) \in \{x_2, x_7\}\}$ 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).

[2]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).

A *contract* is a finite set $\mathcal{C}$ of clauses such that $\bigcap_{C \in \mathcal{C}} C$ is a singleton; let $g_{\mathcal{C}} : A \to X$ denote the sole member of this intersection. The interpretation of $\mathcal{C}$ is that it is a list of statements describing various contingencies of a mechanism, much like a "real world" contract. Each clause $C \in \mathcal{C}$ indicates that $g_{\mathcal{C}} \in C$, so that $\mathcal{C}$ is essentially a set of signals rich enough to pin down a specific mechanism. Formalizing a contract as a set (rather than a sequence) of clauses is without loss of generality because the agent's cognitive process will not depend on the order in which clauses are presented.

To summarize, clauses express information about mechanisms, and contracts are sets of clauses that, combined, are sufficiently informative to pin down a single mechanism. Note that, in this paper, the terms "mechanism" and "contract" are not synonymous: a mechanism is a function $g : A \to X$, while a contract is a particular way of framing or expressing a mechanism—namely, as a set of clauses. Framing mechanisms this way will affect the agent's perception of the underlying mechanism but, in the event of a dispute, a sophisticated third party can verify that $g_{\mathcal{C}}$ is the unique mechanism defined by a contract $\mathcal{C}$.

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 or not to participate in the mechanism. If he does not participate, he retains his outside option. If he participates and takes action $a \in A$, he receives outcome $g_{\mathcal{C}}(a)$, the outcome actually prescribed by $\mathcal{C}$.

2.3 The Agent's Cognitive Process

The idea of the agent's cognitive procedure is that he transitions between different belief states as he processes clauses from a contract. He has a limited ability to combine multiple clauses at once—that is, to perform long chains of reasoning—and this hinders his ability to transition to finer (more informative) states.

Formally, a *belief* is a nonempty-valued correspondence $b : A \rightrightarrows X$. A belief b may be represented by the set $B^b := \{g : A \to X \mid \forall a \; g(a) \in b(a)\}$ 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* and has an associated nonempty-valued correspondence, denoted b^B, such that $B^{b^B} = B$. An agent in state B has narrowed the possibilities for $g_{\mathcal{C}}(a)$ down to the set $b^B(a)$.

There is an integer $K \geq 1$ representing the agent's *deductive ability* (or *working memory*).

An agent of ability K can combine up to K clauses at a time in order to transition between belief states, starting from the state $B = G$. For any finite set S, let $|S|$ denote the cardinality of S.

Definition 1. Let $\mathcal{C}$ be a contract and $K \geq 1$.

1. A *K-valid transition* is a triple $(B, B', \mathcal{C}')$ where $B, B' \in \mathcal{B}$, $\mathcal{C}' \subseteq \mathcal{C}$ is nonempty, $|\mathcal{C}'| \leq K$, and
$$B \cap \left(\bigcap_{C \in \mathcal{C}'} C \right) \subseteq B'$$
The notation $B \xrightarrow{\mathcal{C}'} B'$ indicates the transition $(B, B', \mathcal{C}')$.

2. A state $B \in \mathcal{B}$ is *K-reachable* if there is a sequence of K-valid transitions
$$G = B^0 \xrightarrow{\mathcal{C}^1} B^1 \xrightarrow{\mathcal{C}^2} B^2 \xrightarrow{\mathcal{C}^3} \ldots \xrightarrow{\mathcal{C}^n} B^n = B$$

The interpretation of $B \xrightarrow{\mathcal{C}'} B'$ is that if an agent with deductive ability K is in state B, then he has the ability to transition to state B'. Specifically, he can intersect the clauses of $\mathcal{C}'$ and combine them with his current knowledge, B, to deduce that $g_{\mathcal{C}} \in B'$. This relies on the fact that $|\mathcal{C}'| \leq K$; otherwise, he would lack sufficient working memory to compute the intersection $\bigcap_{C \in \mathcal{C}'} C$. If a state B is K-reachable, then an agent of ability K who begins with no knowledge of $g_{\mathcal{C}}$ can deduce, through a series of K-valid transitions, that $g_{\mathcal{C}} \in B$.

Lemma 1. *If $\mathcal{C}$ is a contract and $K \geq 1$, then there is a unique K-reachable state $B^* \in \mathcal{B}$ such that $B^* \subseteq B$ for all K-reachable states B.*

Lemma 1 states that for every contract $\mathcal{C}$, there exists a finest belief state B^* that is K-reachable. This follows from the fact that $\mathcal{B}$ is closed under nonempty intersections: if $B, B' \in \mathcal{B}$ and $B \cap B' \neq \emptyset$, then $B \cap B' \in \mathcal{B}$. Hence, the desired B^* is simply the intersection of all K-reachable states. For a proof, please see the appendix.

Given Lemma 1, the following is well-defined:

Definition 2 (Induced Belief)**.** The *induced belief state* for an agent of ability K under contract $\mathcal{C}$ is the unique K-reachable state $B_{K,\mathcal{C}} \in \mathcal{B}$ such that $B_{K,\mathcal{C}} \subseteq B$ for all K-reachable $B \in \mathcal{B}$. Let $b_{K,\mathcal{C}}$ denote the associated correspondence; this is the *induced belief*.

This definition says that the agent arrives at the finest possible approximation to $g_{\mathcal{C}}$ given his deductive ability K. Intuitively, the agent repeatedly combines clauses and makes transitions until he gets stuck in a state where further refinement of his beliefs requires a large

leap of logic (the combination of more than K clauses of $\mathcal{C}$). Lemma 1 ensures that there is only one such terminal belief state, despite the many different sequences of transitions that the agent may perform. The definition asserts that he reaches $B_{K,\mathcal{C}}$ but makes no claim about the exact sequence of transitions made along the way. In fact, the process is *path independent* in the sense that it does not matter in what order transitions occur; there is no possibility of getting stuck in a state other than $B_{K,\mathcal{C}}$.

Upon forming beliefs $b_{K,\mathcal{C}}$, an agent of 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 criterion when evaluating the set of outcomes $b_{K,\mathcal{C}}(a)$ that he considers possible at actions $a \in A$. Effectively, his cognitive limitation leads him to believe that the contract is ambiguous, and he is averse to this perceived ambiguity. This is an extreme degree of ambiguity aversion, but many insights generated by the model hold under alternative assumptions; see section 4.3.

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.

While ambiguity aversion appears to be a common phenomenon, there is not widespread agreement as to why individuals display this characteristic. One intriguing possibility that seems particularly well-suited to the present setting is the idea of *deceit aversion*. In the implementation model, deceit aversion (and, hence, ambiguity aversion) reflects the attitude of an agent who is aware of his cognitive limitation and skeptical of the principal's motives: the fact that the agent cannot pin down the mechanism raises suspicion that the principal is trying to deceive him. Only a worst-case criterion is guaranteed to protect cognitively constrained agents from bad outcomes (those dominated by their outside options). Hence, in the presence of potential manipulators, this form of ambiguity aversion may be an advantageous heuristic for cognitively constrained individuals.

2.4 Comments on the Agent's Procedure

The purpose of this section is to elaborate on the interpretation of, and motivation for, the model of bounded rationality described above. Readers who prefer to skip to the implementation results will find them in section 3.

In general, this model of bounded rationality is meant to capture some aspects of what it means for a problem to be difficult (for a human) to solve. To achieve this, I assume that the agent engages in a particular procedure for processing information. As illustrated by the Sudoku example in the introduction, a complex puzzle is one where the agent lacks the sophistication required to reach the solution despite understanding all of the rules. In the implementation model, the rules are clauses of $\mathcal{C}$ and the solution is the underlying mechanism $g_{\mathcal{C}}$. Simple puzzles (contracts) are those that can be resolved in small steps, never requiring large "leaps of logic" involving the combination of many rules (clauses) at once. In a complex puzzle (contract), an agent of ability K may get stuck in a belief state B where the only way to transition to a strictly finer B' is to combine more than K rules (clauses).

Despite appearances, the assumption that the agent understands all clauses of a contract does *not* mean that he understands all of the (logically equivalent) ways of expressing a clause in formal or natural language. Rather, this is an assumption about the expressive power of the principal: she has the ability to convey properties of $g_{\mathcal{C}}$ in a way that all agents understand. To see the difference, consider four statements: (i) "$3x + y = 11$", (ii) "$2x - y = 4$", (iii) "$x = 3$ and $y = 2$", and (iv) "$3x + y = 11$ and $2x - y = 4$". (To relate this back to the implementation problem, think of x and y as the consequences of two different actions). Clearly, (iii) and (iv) are logically equivalent. However, it is conceivable that a person understands (iii) but not (iv) because (iv) requires some amount of calculation to pin down x and y. In my model, (iv) is treated like a set of clauses $\mathcal{C} = \{(i), (ii)\}$ that pins down the values of x and y. The agent will correctly infer these values from $\mathcal{C}' = \{(iii)\}$, but not necessarily from $\mathcal{C}$. If the principal wishes to express $x = 3$ and $y = 2$ in a way that the agent understands, she will choose $\mathcal{C}'$; if she is interested in exploiting the agent's cognitive ability (while still committing to $x = 3$ and $y = 2$), she might choose $\mathcal{C}$ instead. The fact that a collection of clauses is logically equivalent to a single clause does not conflict with the interpretation of the model, because the model does not assume that the agent understands all of the different ways that a single clause might be expressed.

I assume that every set C of mechanisms can be clearly expressed because, for now, I am not interested in how the mechanics of restricted language may impact the principal's incentives to design complex contracts (but see the discussion of Result 4 below). Any

assumption about language (ie, which sets C can be clearly expressed) is necessarily arbitrary and, under any such assumption, the issue of how an agent processes multiple pieces of information remains relevant. I have chosen to emphasize the latter issue, and feel that the assumption of unrestricted language reinforces a broader point: the principal *chooses* to frame mechanisms in complex, manipulative ways despite having the ability to communicate them clearly. Consequently, my model captures the idea that from the principal's perspective, a well-designed contract (like a well-designed puzzle) has clear and simple rules (clauses) but is nonetheless difficult to resolve.

Still, one might wonder what happens under alternative assumptions. What if the agent does not understand some clauses, or is too impatient to perform all of the transitions needed to arrive at the finest possible beliefs? It turns out the optimal contract derived in the baseline model is quite robust to many considerations of this type; see the discussion of Result 3 below.

3 Implementation via Complex Contracts

To simplify the exposition, I restrict attention to direct mechanisms (those where $A = \Theta$). The techniques developed here can easily be adapted to the case of arbitrary action spaces—see section 4.2. Throughout, I assume $\overline{x} = (\overline{x}_\theta)_{\theta \in \Theta}$ is a fixed profile of outside options.

3.1 Main Results

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 $f(\theta) \succsim_\theta \overline{x}_\theta$. The primary goal of this section is to characterize when and how nontrivial functions can be implemented for boundedly rational agents.

For an agent of ability K, the principal seeks to design a contract $\mathcal{C}$ that *K-implements* the objective function f:

Definition 3 (K-Implementation)**.** Fix an integer $K \geq 1$. A contract $\mathcal{C}$ *K-implements* the function $f : \Theta \to X$ if the following conditions are met:

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

A function f is *implementable* if it is K-implementable for some K.

The Incentive Compatibility (IC) condition requires the contract to induce beliefs making truthful reporting a best response whenever the agent chooses to participate.[3] The Individual Rationality (IR) condition requires the agent to expect (via the worst-case criterion) an outcome at least as good as his outside option if he chooses to participate and respond truthfully. The final requirement states that the outcome the agent actually receives after reporting θ is $f(\theta)$.

Definition 3 does not require truthful reporting to be a unique best response. For an analysis of strict implementation—implementation where truthful reporting is the unique optimal response—see section 4.1.

The analysis is organized into four results. The first is:

Result 1: Complexity expands the set of implementable functions

In particular, a function is implementable if and only if it satisfies a simple dominance condition involving the profile of outside options. Trivial functions satisfy the condition and, typically, so do many non-trivial 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 $\succsim_\theta$. The condition is defined as follows:

Definition 4 (IR Dominance)**.** A function $f : \Theta \to X$ *IR-Dominates* $\overline{x} = (\overline{x}_\theta)_{\theta\in\Theta}$ 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 weakly prefer $f(\theta)$ over every x such that $L_\theta(x) \subseteq L_{\theta'}(\overline{x}_{\theta'})$ for some θ'. To see why this is a necessary condition for K-implementability, suppose a contract $\mathcal{C}$ K-implements f. Then the induced beliefs $b_{K,\mathcal{C}}$ must exclude all outcomes $y \in L_{\theta'}(\overline{x}_{\theta'})$ from the set $b_{K,\mathcal{C}}(\theta')$, or else the IR condition is violated for type θ'. Therefore, if $L_\theta(x) \subseteq L_{\theta'}(\overline{x}_{\theta'})$, then type θ may prefer to misreport as type θ' unless $b_{K,\mathcal{C}}(\theta)$ eliminates a superset of $L_\theta(x)$. Since $f(\theta) = g_{\mathcal{C}}(\theta) \in b_{K,\mathcal{C}}(\theta)$, this forces $f(\theta) \succsim_\theta x$. Proposition 1 establishes that IR dominance is also a sufficient condition for implementability:

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

This characterization allows one to determine whether or not a function is implementable by studying only the function, the preferences, and the outside options; no details concerning the agent's bounded rationality are required. However, Proposition 1 makes no claim about

[3]The focus on separating schemes (rather than allowing types to pool on a particular response if they are to receive the same outcome) is for expository convenience; the techniques developed here can easily be adapted to consider pooling schemes as well.

which levels of K are acceptable and does not rule out the possibility that different functions are implementable for different ranges of K.

Result 2: The optimal contract is robust to variation in K

In fact, the next proposition shows that it is without loss of generality to consider a robust form of implementation where a contract K-implements a function for all K up to some bound. Rather than designing different contracts for different abilities K, the principal can design a single contract that achieves implementation for all K below a particular bound. This bound depends on the profile $\overline{x}$ of outside options, but not on the objective function f. The implementing contract takes the form of a *complex contract*, defined next.

Definition 5 (Complex Contract)**.** Let $f \in D(\overline{x})$. The *complex contract for f* is given by

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

Each clause $C \in \mathcal{C}_f$ allows the agent to deduce that $g_{\mathcal{C}}$ is IR Dominant, but provides only slightly more information: it eliminates precisely one mechanism from $D(\overline{x})$. This maximizes the number of clauses that must be combined in order to pin down $g_{\mathcal{C}}$, conditional on the IR Dominance property of $g_{\mathcal{C}}$ being deducible for all cognitive abilities K.

The main result of this paper establishes that the principal need only consider complex contracts:

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

Proposition 2 says that the complex contract $\mathcal{C}_f$ is always an optimal contract from the principal's perspective: if some other contract K-implements f, then so does $\mathcal{C}_f$. Hence, the principal does not require—and cannot benefit from—any knowledge of the agent's ability K: she can do no better than to present the complex contract and hope that the agent's ability K is not sufficiently large to deduce the true mechanism $g_{\mathcal{C}_f}$. If K is too large (and f is nontrivial), no other contract would achieve implementation anyway.

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

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 $\overline{K}(\overline{x})$ is, in part, determined by the choice of action space $A = \Theta$. In section 4.2, I show that $\overline{K}(\overline{x})$ can be made arbitrarily large (without affecting the set of implementable functions) by enlarging the action space. Hence, a principal with

sufficient expressive power and little to no cost of producing long contracts may prefer to inflate the action space.

As the notation suggests, the set of implementable functions $D(\overline{x})$ and the bound $\overline{K}(\overline{x})$ vary with the profile $\overline{x}$ of outside options:

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 IR Dominant functions shrinks and $\overline{K}$ decreases as outside options become more attractive for all types. 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. This case still requires the full proof, sketched in section 3.3, to establish both implementability as well as the optimality of complex contracts.

Result 3: The agent only needs to process one clause

The model assumes that the agent continues to process clauses and transition to new belief states until he gets stuck in a state where further refinement of his beliefs requires the combination of more than K clauses. In theory, reaching such a terminal state may require the agent to perform many transitions, especially if K is small. What if the agent gets tired or impatient? If he terminates the procedure prematurely, will he still hold incentive compatible beliefs?

A key feature of $\mathcal{C}_f$ is that every individual $C \in \mathcal{C}_f$ allows the agent to deduce that $g_{\mathcal{C}} \in D(\overline{x})$. Since $D(\overline{x}) \in \mathcal{B}$, this means $G \xrightarrow{\{C\}} D(\overline{x})$ is a K-valid transition for all K. In other words, any randomly chosen clause of $\mathcal{C}_f$ will induce incentive-compatible beliefs. In fact, once an agent arrives in belief state $D(\overline{x})$, he can only transition to finer beliefs if he has ability $K \geq \overline{K}(\overline{x})$—in which case he is sophisticated enough to deduce the true mechanism. Thus, $\mathcal{C}_f$ has the property that it is very easy for agents to end up with incentive compatible beliefs, but very difficult for them to reach finer beliefs.

This is reassuring because it suggests that $\mathcal{C}_f$ is either optimal or nearly optimal under plausible variations on the agent's cognitive procedure. For example, suppose that the agent processes clauses stochastically: a clause C is processed with some probability π_C. This randomness could be due to imperfect communication by the principal (ie, despite the principal's best efforts, the agent might fail to understand a clause), or the agent may simply fail to pay full attention to all clauses (eg, the contract is too long and the agent randomly selects a subset of clauses). Since only one clause of $\mathcal{C}_f$ needs to be processed for the agent to arrive at incentive compatible beliefs, but many need to be combined at once to refine those beliefs, one would expect $\mathcal{C}_f$ to have a high success rate. In other words, $\mathcal{C}_f$ is robust

not only to variation in K, but also to nearby models of cognition that are perhaps more realistic (but also more complicated).

Result 4: Complex contracts can be mutually beneficial

The analysis so far has focused on the case of nontrivial objective functions. When the objective function is nontrivial, the preferences of the agent conflict with the goals of the principal, and complex contracts arise out of the principal's incentive to exploit the agent's limited cognitive ability.

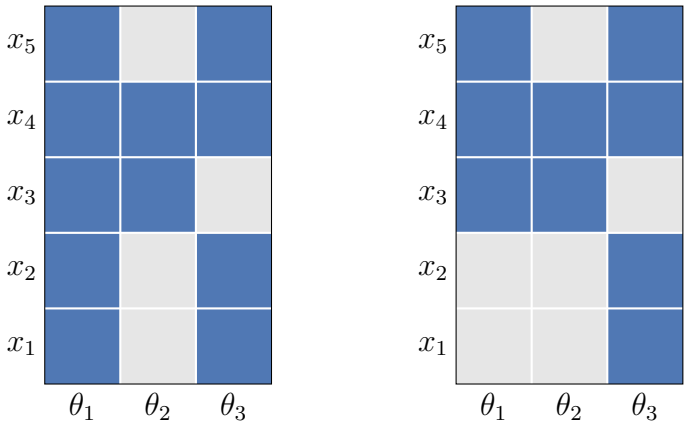
The preceding result, however, suggests that complex contracts can be a useful tool even in the absence of such conflicts. For example, a company offering a menu of medical insurance plans to its employees may have the same objective as the employees (choose the plan that best fits the needs of the particular employee), but the plans themselves may be detailed and nuanced. Suppose that the action set A is the list of available insurance plans and that the outcome set X is some larger domain of plans—essentially, all combinations of attributes that a plan could conceivably have. The principal (company) would like to express the identity function $g : A \to X$ given by $g(a) = a$; in other words, the principal would like the agent to fully understand the available options so that the agent can identify and then choose his most-preferred plan.

If the principal and agent are constrained by language—that is, if the principal lacks the ability to clearly express the function g in a way that the agent understands—then the principal must express g as a collection of clauses. In other words, the lack of a rich common language *forces* a complex description of g. If the agent does not successfully process these clauses, he may arrive at beliefs about g that result in a suboptimal choice.

The previous result suggests a useful guideline for the principal: try to formulate a set of clauses such that successful combination of any subset of clauses induces incentive compatible beliefs. It is not enough to simply choose *some* contract $\mathcal{C}$ that defines g, even if all of the clauses are understandable; if the agent is boundedly rational, an ideal contract guides the agent toward his most-preferred response even if he fails to fully comprehend the underlying mechanism. The complex contract $\mathcal{C}_f$ may be a useful baseline because, in this setting, it achieves implementation as long as the agent correctly processes at least one clause (and processing additional clauses will keep the agent in an incentive compatible belief state).

3.2 Example

To illustrate the main results, consider the following model. There are three types θ_1, θ_2, θ_3 and five outcomes x_i $(i = 1, \ldots, 5)$. The outside options are $\overline{x}_{\theta_1} = x_1$, $\overline{x}_{\theta_2} = x_4$, and



(a) $L_\theta(\overline{x}_\theta)$ removed (b) $L^*_\theta(\overline{x})$ removed

Figure 2: Constructing b^*

$\overline{x}_{\theta_3} = x_2$. Preferences are given by the following table (ordering best to worst for each $\succsim_i$):

$\succsim_1$	x_5	x_4	x_3	x_2	x_1
$\succsim_2$	x_3	x_4	x_5	x_2	x_1
$\succsim_3$	x_4	x_5	x_1	x_2	x_3

To compute the set of implementable functions, it is enough to find the largest correspondence b inducing all agent types to participate and report truthfully. Clearly, this requires $b(\theta)$ to exclude every outcome in $L_\theta(\overline{x}_\theta)$, or else type θ prefers $\overline{x}_\theta$ over reporting θ. As Figure 2a shows, however, it is not enough to only remove these outcomes since the resulting correspondence need not satisfy the IC condition. In particular, type θ_1 would prefer to misreport as θ_2 because (by the worst-case criterion) he expects outcome x_1 from reporting θ_1 and $x_3 \succ_{\theta_1} x_1$ from reporting θ_2.

The minimal change needed to make θ_1 an optimal response is to remove x_1 and x_2 as possible outcomes from reporting θ_1; call this new lower contour set $L^*_{\theta_1}(\overline{x})$. This makes type θ_1 indifferent between reporting θ_1 and θ_2, while also ensuring that truthful reporting is a best response for types θ_2 and θ_3; see Figure 2b. This is the sought-after correspondence b^*.

For each f such that $f(\theta) \in b^*(\theta)$, the complex contract $\mathcal{C}_f = \{B^{b^*} \backslash \{g\} : g \in B^{b^*} \text{ and } g \neq f\}$ K-implements f for all $K < \overline{K}(\overline{x})$. How is $\overline{K}(\overline{x})$ computed? Suppose an agent of ability K arrives in state B^{b^*} (only one clause of $\mathcal{C}_f$ must be processed for this to occur). If the agent is able to transition to a proper subset of B^{b^*}, then he must be able to eliminate some point (θ, x) from the correspondence b^*. Since each $C \in \mathcal{C}_f$ eliminates exactly one function from B^{b^*}, this requires the combination of $K_{\theta,x}$ clauses from $\mathcal{C}_f$, where $K_{\theta,x}$ is the number of functions in B^{b^*} passing through the point (θ, x). Hence, $K \geq K_{\theta,x}$.

Notice that $K_{\theta,x}$ does not depend on the choice of x; the only requirement is that $x \in b^*(\theta)$ and $x \neq f(\theta)$. Hence, for any $y \in b^*(\theta) \backslash \{f(\theta), x\}$, the fact that $K \geq K_{\theta,x} = K_{\theta,y}$ implies

that the agent will be able to eliminate the point (θ, y) as well. Continuing in this fashion, he eventually pins down $f(\theta)$.

The resulting belief state B contains far few functions than B^{b^*} because $g(\theta) = f(\theta)$ for all $g \in B$. In fact, for any $\theta' \neq \theta$ and any $x' \in b(\theta')\backslash\{f(\theta')\}$, there are exactly

$$\prod_{\theta'' \neq \theta'} |b(\theta'')| = \prod_{\theta'' \neq \theta', \theta} |b^*(\theta'')| \leq \prod_{\theta'' \neq \theta} |b^*(\theta'')| = K_{\theta,x}$$

functions in B passing through (θ', x'). Hence, the agent will be able to eliminate (θ', x') and, by a similar argument to that above, pin down $f(\theta')$. Repeating this logic, he must eventually pin down the function f.

Thus, if the agent is able to reach any proper subset of B^{b^*}, then he must eventually deduce f. If f is nontrivial, this means f will not be implemented. Hence, to guarantee implementation, it must be the case that

$$K < \min_{\theta \in \Theta} \prod_{\theta' \neq \theta} |b^*(\theta')|$$

so that the agent is never able to eliminate any points from $b^*(\theta)$ for any θ. This bound is the desired $\overline{K}(\overline{x})$.

3.3 Sketch of the Proof

To implement a function f, the principal must formulate a contract $\mathcal{C}$ such that (i) $g_{\mathcal{C}} = f$ and (ii) the induced belief $b_{K,\mathcal{C}}$ satisfies the IR and IC constraints. Since $g_{\mathcal{C}} = f$, it follows that $g_{\mathcal{C}}(\theta) \in b_{K,\mathcal{C}}(\theta)$ for all θ; that is, $f \in B_{K,\mathcal{C}}$ (the agent does not eliminate f itself as a candidate for $g_{\mathcal{C}}$).

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 establishes three results. First, it shows that $D(\overline{x}) \in \mathcal{B}$; that is, there is a correspondence $b^* : A \rightrightarrows X$ such that $D(\overline{x}) = \{g \in G : \forall \theta \in \Theta,\ g(\theta) \in b^*(\theta)\}$. Second, it shows that (as a belief state) $D(\overline{x})$ satisfies the IC and IR constraints. Third, it shows that b^* is the largest belief correspondence satisfying the IC and IR constraints; 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, then $f \in B_{K,\mathcal{C}} \subseteq D(\overline{x})$, so that IR-dominance is a necessary condition for implementability.

Given Lemma 2, the objective is to show that every IR-dominant function is implementable and that an IR-dominant function is K-implementable (for some K) if and only if it is K-implementable by the complex contract $\mathcal{C}_f$. For the remainder of this section, let $f \in D(\overline{x})$ denote an IR-dominant function.

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

Lemma 3 says that if $\mathcal{C}$ induces beliefs $B_{K,\mathcal{C}}$ satisfying the IC and IR constraints, then an agent of ability K forms coarser beliefs under the complex contract $\mathcal{C}_f$. The intuition for this result is as follows. Since $\mathcal{C}$ K-implements f, it follows that $B_{K,\mathcal{C}} \subseteq D(\overline{x})$ (Lemma 2). Thus, the state $D(\overline{x})$ is K-reachable under $\mathcal{C}$. Clearly, $D(\overline{x})$ is also K-reachable under $\mathcal{C}_f$ (each $C \in \mathcal{C}_f$ is a subset of $D(\overline{x})$). Since each $C \in \mathcal{C}_f$ is of the form $C = D(\overline{x})\backslash\{g\}$ (where $g \in D(\overline{x})$), the contract $\mathcal{C}_f$ maximizes the number of clauses that must be combined in order to transition between subsets of $D(\overline{x})$. That is, if $B, B' \in \mathcal{B}$ and $B' \subsetneq B \subseteq D(\overline{x})$, then more clauses of $\mathcal{C}_f$ (compared to $\mathcal{C}$) must be combined in order to transition from B to B'. Hence, $\mathcal{C}_f$ yields (weakly) coarser beliefs for all K.

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

The idea behind Lemma 4 is as follows. The state $D(\overline{x})$ is K-reachable for all K. But every $B \in \mathcal{B}$ such that $B \subsetneq D(\overline{x})$ represents a belief correspondence b^B that is a proper sub-correspondence of b^*. Hence, there is a pair (θ, x) such that $x \in b^*(\theta)$ but $x \notin b^B(\theta)$. In order to transition to such a B from state $D(\overline{x})$, every function $g \in D(\overline{x})$ passing through (θ, x) must be eliminated. This requires the successful combination of $\prod_{\theta' \neq \theta} |b^*(\theta')|$ clauses of $\mathcal{C}_f$ because each $C \in \mathcal{C}_f$ eliminates exactly one function in $D(\overline{x})$. Hence, such a transition requires $\prod_{\theta' \neq \theta} |b^*(\theta')| \leq K$. Note that this bound does not depend on the chosen x. Hence, after eliminating (θ, x), such an agent will be able to eliminate (θ, x') for any $x' \in b^*(\theta)\backslash\{x\}$ such that $x' \neq f(\theta)$. Continuing in this way, the state $B' = \{g \in D(\overline{x}) : g(\theta) = f(\theta)\}$ is K-reachable. From B', he will now be able to eliminate a point (θ', x'') where $\theta' \neq \theta$ and $x'' \in b^*(\theta')\backslash\{f(\theta)\}$ because the number of functions in B' passing through (θ', x'') is $\prod_{\hat{\theta} \neq \theta, \theta'} |b^*(\hat{\theta})| \leq K$. Repeating this argument, the state $\{f\} \in \mathcal{B}$ is K-reachable.

Armed with these Lemmas, the remainder of the proof is fairly straightforward. First, Lemma 4 implies that if

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

then $B_{K,\mathcal{C}_f} = D(\overline{x})$, and that if $K > \overline{K}(\overline{x})$, then $B_{K,\mathcal{C}_f} = \{f\}$. This implies that every IR-dominant function is K-implementable by $\mathcal{C}_f$ for some K (namely $K = 1$), proving Proposition 1.[4]

[4]In some cases, $\overline{K}(\overline{x}) = 1$ and a slightly modified argument is needed; see the appendix.

Next, 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 $K \leq \overline{K}(\overline{x})$, then $B_{K,\mathcal{C}_f} = D(\overline{x})$ and therefore $\mathcal{C}_f$ K-implements f for all $K' \leq \overline{K}(\overline{x})$. If $K > \overline{K}(\overline{x})$, 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. This establishes Proposition 2.

4 Extensions

4.1 Strict Implementation

The definition of K-implementation does not require the agent to strictly prefer reporting his true type θ. If a contract induces beliefs making the agent indifferent between multiple responses, then he may not truthfully report his type unless he suffers a small cost of lying or, more generally, is "white lie averse".[5]

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

Definition 6 (Strict K-Implementation)**.** Fix an integer $K \geq 1$. A contract $\mathcal{C}$ *strictly K-implements* the function $f : \Theta \to X$ if the following conditions are met:

1. (Strict Incentive Compatibility). For all $\theta' \neq \theta \in \Theta$, $U_\theta(\theta, K, \mathcal{C}) > U_\theta(\theta', K, \mathcal{C})$.

2. (Individual Rationality). For all $\theta \in \Theta$, $U_\theta(\theta, K, \mathcal{C}) \geq u_\theta(\overline{x}_\theta)$.

3. For all $\theta \in \Theta$, $g_{\mathcal{C}}(\theta) = f(\theta)$.

A function f is *strictly implementable* if it is strictly K-implementable for some K.

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 7 (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 $\theta = \theta'$.

[5] 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).

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) that contain $\overline{x}_\theta$ but not each other, and such that type θ weakly prefers $f(\theta)$ over any $y \in L_\theta(x^*_\theta)$. 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, $D^*(\overline{x})$ is a member of $\mathcal{B}$ and the techniques developed to analyze K-implementation apply in the strict case as well.

In particular, for any $f \in D^*(\overline{x})$, let the *strict complex contract* for f be defined by:

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

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

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 $\overline{K}^*(\overline{x}) \leq \overline{K}(\overline{x})$. That is, 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.

4.2 Arbitrary Action Sets

So far, I have only considered the case $A = \Theta$. 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 $D(\overline{x}) \in \mathcal{B}$. 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

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

It is easy to see that $\mathcal{C}_f^A$ K-implements f for all K below a bound $\overline{K}^A(\overline{x})$ that is increasing in the cardinality of $A\backslash\Theta$, and that only IR Dominant functions are implementable. In particular,

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

This suggests that the principal may wish to inflate A indefinitely, thereby achieving implementation for any K she desires. In practice, the principal may be constrained by costs associated with generating larger contracts as well as language needed to distinguish elements in a larger action space.

4.3 Ambiguity Attitude

Although I have assumed the agent uses a worst-case criterion to resolve the (perceived) ambiguity in his beliefs $b_{K,\mathcal{C}}$, many of the insights generated in this case also hold under alternative assumptions. This is so because the agent's procedure for forming beliefs is independent of how he evaluates actions $a \in A$ conditional on those beliefs: his ambiguity attitude has no bearing on whether or not a state B is reachable. Therefore, a similar two-step approach can be used to analyze the model under alternative assumptions:

1. Characterize which belief correspondences b satisfy appropriate IC and IR constraints
2. For any such b and objective function $f \in B^b$, take a complex contract of the form

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

Such a $\mathcal{C}_f^b$ will K-implement f for all $K < \min_{\theta\in\Theta} \prod_{\theta'\neq\theta} |b(\theta')|$. Hence, alternative assumptions affect the set of implementable functions, but not the result that this set expands with boundedly rational agents, nor the result that the principal can restrict attention to complex contracts that achieve robust implementation.

5 Conclusion

In this paper, I have shown how a sophisticated designer can manipulate a cognitively constrained agent by carefully selecting a set of rules (a contract) to be processed by the agent. The principal in my model has the ability to clearly specify any mechanism in a way that the agent understands, but chooses a complex framing of it in order to achieve goals that, under full rationality, cannot be implemented. Although agents may vary in their degree

of cognitive sophistication, the principal will (without loss of generality) select a complex contract achieving implementation for as wide a range of cognitive abilities as possible—the principal does not require any information about the agent's ability in order to formulate the optimal framing.

Central to the analysis is the model of bounded rationality. In this model, the agent processes clauses of the contract in order to transition between belief states. Belief states represent knowledge that the agent can hold and reason about, and he can only combine his current beliefs with up to K clauses at a time when transitioning to new states. Hence, in this framework, a set of clauses (a contract) is more complex if a higher ability K is required to perform a sequence of transitions pinning down the true mechanism. In contrast, a simple contract can be understood by proceeding in small steps, never requiring a "leap of logic" (the combination of many clauses) in order to transition to finer beliefs.

Clearly, this model of bounded rationality need not be confined to the domain of implementation theory. It can be reformulated, for example, in a standard state space setting where Ω is a set of states and $\mathcal{B}$ is a family of subsets of Ω (belief states) closed under nonempty intersections. A frame for an event $E \subseteq \Omega$ is a family $\mathcal{F}$ of subsets of Ω such that $\bigcap_{F \in \mathcal{F}} F = E$; in other words, E is framed as a set of signals F that jointly pin down E. For any $K \geq 1$, the cognitive procedure for processing $\mathcal{F}$ can clearly be adapted from the agent's procedure in this paper, providing an intuitive theory of framing (complexity) in information processing.

References

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 (2014). Mechanism design with bounded depth of reasoning and small modeling mistakes. *Available at SSRN 2460019*.

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.

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

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.

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 (2013). Contracts with framing. *Working paper*.

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.

A Proofs for Sections 2 and 3

Lemmas 5 and 6 state some basic results about $\mathcal{B}$, K-valid transitions, and K-reachability that will be used repeatedly in subsequent arguments. The proofs of these lemmas are elementary and therefore omitted.

Lemma 5. *The family $\mathcal{B}$ is closed under nonempty intersections: if $B, B' \in \mathcal{B}$ and $B \cap B' \neq \emptyset$, then $B \cap B' \in \mathcal{B}$. Since $\mathcal{B}$ is finite, it follows that $\mathcal{B}$ is closed under arbitrary nonempty intersections.*

Lemma 6. *The following holds for all contracts $\mathcal{C}$ and all $K \geq 1$:*

(i) If $B \xrightarrow{\mathcal{C}'} B'$ is K-valid and $B' \subseteq B'' \in \mathcal{B}$, then $B \xrightarrow{\mathcal{C}'} B''$ is K-valid.

(ii) If B is K-reachable and $B \xrightarrow{\mathcal{C}'} B'$ is K-valid, then B' is K-reachable.

(iii) If B and B' are K-reachable, then $B \cap B' \neq \emptyset$. Hence, $B \cap B' \in \mathcal{B}$.

(iv) If $K' \geq K$, then $B_{K',\mathcal{C}} \subseteq B_{K,\mathcal{C}}$.

Lemma 1. *If $\mathcal{C}$ is a contract and $K \geq 1$, then there is a unique $B^* \in \mathcal{B}$ such that $B^* \subseteq B$ for all K-reachable states B.*

Proof. It is enough to show that if B and B' are K-reachable, then $B \cap B'$ is K-reachable. The lemma follows by taking B^* to be the intersection of all K-reachable states.

So, suppose

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

Then

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

Since

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

it follows that $B = B \cap B^0 \xrightarrow{\mathcal{C}^1} B \cap B^1$ is a K-valid transition. Proceeding inductively, suppose $i < n$ and that $B = B \cap B^0 \xrightarrow{\mathcal{C}^1} B \cap B^1 \xrightarrow{\mathcal{C}^2} B \cap B^1 \cap B^2 \xrightarrow{\mathcal{C}^3} \ldots \xrightarrow{\mathcal{C}^i} B \cap B^0 \cap \ldots \cap B^i$ is a sequence of K-valid transitions. Then

$$B \cap B^0 \cap \ldots \cap B^i \cap \left(\bigcap_{C \in \mathcal{C}^{i+1}} C \right) \subseteq B^i \cap \left(\bigcap_{C \in \mathcal{C}^{i+1}} C \right) \subseteq B^{i+1}$$

Thus

$$B \cap B^0 \cap \ldots \cap B^i \cap \left(\bigcap_{C \in \mathcal{C}^{i+1}} C \right) \subseteq B \cap B^0 \cap \ldots \cap B^i \cap B^{i+1}$$

and $B \cap B^0 \cap \ldots \cap B^i \xrightarrow{\mathcal{C}^{i+1}} B \cap B^0 \cap \ldots \cap B^{i+1}$ is a K-valid transition. Hence, $B \cap B^0 \cap \ldots \cap B^n$ is K-reachable. Since $B^n = B'$, it follows that $B \cap B^0 \cap \ldots \cap B^n \subseteq B \cap B'$, and so $B \cap B'$ is K-reachable. □

A.1 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. Then any two sets $L_\theta(Y)$, $L_\theta(Y')$ are ordered by set inclusion. Take L^*_θ to be the largest set $L_\theta(Y)$ among all sets $Y = L_{\theta'}(\overline{x}_{\theta'})$ ($\theta' \in \Theta'$), and let $b^*(\theta) := X \backslash L^*_\theta$. The following Lemma expands upon Lemma 2:

Lemma 7. *The set $D(\overline{x})$ of all IR-dominant functions satisfies the following:*

(i) $D(\overline{x})$ consists of all f such that $f(\theta) \notin L^*_\theta(\overline{x})$. Hence, $D(\overline{x}) \in \mathcal{B}$ with associated correspondence b^*.

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

(iii) Every belief state $B_{K,\mathcal{C}}$ 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 $B = \{f : \Theta \to X \mid \forall \theta \ f(\theta) \notin L^*_\theta(\overline{x})\}$. 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 $L^*_\theta(\overline{x}) = L_\theta(Y)$ where $Y = L_{\theta^*}(\overline{x}_{\theta^*})$. Since $L^*_\theta(\overline{x})$ is a strict lower contour, there is an $x^* \in X$ such that $L^*_\theta(\overline{x}) = L_\theta(x^*)$. Then $L_{\theta^*}(\overline{x}_{\theta^*}) \supseteq L_\theta(x^*)$, so that by IR dominance $f(\theta) \succsim_\theta x^*$. Since $x^* \succ_\theta x$ for all $x \in L_\theta(x^*) = L^*_\theta(\overline{x})$, it follows that $f(\theta) \notin L^*_\theta(\overline{x})$.

For the converse inclusion, suppose $f \in B$ and that $L_{\theta'}(\overline{x}_{\theta'}) \supseteq L_\theta(x)$. We need to show that $f(\theta) \succsim_\theta x$. We have $f(\theta) \notin L^*_\theta(\overline{x})$, and therefore $f(\theta) \succ_\theta x'$ for all $x' \in L^*_\theta(\overline{x})$. In particular, $f(\theta) \succsim_\theta x$ because $L_\theta(x) \subseteq L_\theta(L_{\theta'}(\overline{x}_{\theta'})) \subseteq L^*_\theta(\overline{x})$. Thus, $f \in D(\overline{x})$. □

Proof of (ii). By (i), we may represent $D(\overline{x})$ by the set $B = \{f : \Theta \to X \mid \forall \theta \ f(\theta) \notin L^*_\theta(\overline{x})\}$. Clearly, this set satisfies the IR condition. For the IC condition, suppose toward a contradiction that some type θ strictly prefers to misreport as some $\theta' \neq \theta$ under beliefs B. This implies that $L^*_\theta(\overline{x}) \subsetneq L_\theta(L^*_{\theta'}(\overline{x}))$; that is, $L^*_{\theta'}(\overline{x})$ contains a strictly larger lower contour set of $\succsim_\theta$ than $L^*_\theta(\overline{x})$. Now, there is a θ^* such that $L^*_{\theta'}(\overline{x}) = L_{\theta'}(L_{\theta^*}(\overline{x}_{\theta^*}))$. Then

$L^*_{\theta'}(\overline{x}) \subseteq L_{\theta^*}(\overline{x}_{\theta^*})$, which implies $L_\theta(L^*_{\theta'}(\overline{x})) \subseteq L_\theta(L_{\theta^*}(\overline{x}_{\theta^*}))$. But then $L^*_\theta(\overline{x}) \subsetneq L_\theta(L^*_{\theta'}(\overline{x})) \subseteq L_\theta(L_{\theta^*}(\overline{x}_{\theta^*}))$. This contradicts the fact that $L^*_\theta(\overline{x})$ 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 $B_{K,\mathcal{C}} = B$ 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 $x \in L^*_\theta(\overline{x})$ (because $x \notin b^*(\theta) = X \backslash L^*_\theta(\overline{x})$ 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 $L^*_\theta(\overline{x})$, there exists θ^* such that $L^*_\theta(\overline{x}) = L_\theta(L_{\theta^*}(\overline{x}_{\theta^*}))$. We must have $\theta^* \neq \theta$; otherwise, $L^*_\theta(\overline{x}) = L_\theta(\overline{x}_\theta)$, contradicting the fact that $x \in L^*_\theta(\overline{x}) \backslash L_\theta(\overline{x}_\theta)$.

Next, observe that $y \notin b(\theta^*)$ for all $y \in L_{\theta^*}(\overline{x}_{\theta^*})$ 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 (through the worst-case criterion) an outcome strictly better than x from reporting as type θ^*, because $x \in L^*_\theta(\overline{x}) = L_\theta(L_{\theta^*}(\overline{x}_{\theta^*}))$ and no element of $L_\theta(L_{\theta^*}(\overline{x}_{\theta^a st}))$ (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 unique 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 $g(\theta^*) \succsim_{\theta^*} \overline{x}_{\theta^*}$. Thus, ever $g \in D(\overline{x})$ is trivial. □

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

Proof. Clearly, $D(\overline{x}) \in \mathcal{B}$ is K-reachable under $\mathcal{C}_f$ for all K (simply 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 7, $B_{K,\mathcal{C}} \subseteq D(\overline{x})$ 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 $B \xrightarrow{\widehat{\mathcal{C}}} B'$ is K-valid. This implies that every K-reachable subset of $D(\overline{x})$ under $\mathcal{C}_f$ is K-reachable under $\mathcal{C}$. Since $B_{K,\mathcal{C}_f} \subseteq D(\overline{x})$ and $B_{K,\mathcal{C}} \subseteq D(\overline{x})$ (Lemma 2) and induced beliefs are the intersection of all K-reachable sets, the result follows.

So, suppose $B \xrightarrow{\mathcal{C}'} B'$ is K-valid for some $\mathcal{C}' \subseteq \mathcal{C}_f$. 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

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

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

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

Combined with (1), 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. □

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 demonstrated 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 from $A = \Theta$ to X associated with the set $D(\overline{x})$. 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. Note that $B^i \subseteq D(\overline{x})$ 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 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 $\mathcal{C}' = \{D(\overline{x}) \backslash \{g'\} : g' \in E\}$ where E contains every $g' \in D(\overline{x})$ such that $g'(\theta) = x$. Thus,

$$|\{g \in D(\overline{x}) : g(\theta) = x\}| \leq K \tag{2}$$

Clearly, (2) holds for every choice of $x \in b^*(\theta)$ such that $x \neq g_{\mathcal{C}_f}(\theta)$ because $|\{g \in D(\overline{x}) : g(\theta) = x\}| = \prod_{\theta' \neq \theta} |b^*(\theta')|$.

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 $\hat{B}^i \in \mathcal{B}$ satisfies $b^{\hat{B}^i}(\theta) = b^*(\theta) \backslash \{x_1, \ldots, x_i\}$. The transitions are K-valid because $|\mathcal{C}^{(\theta,x_i)}| = |\{g \in D(\overline{x}) : g(\theta) = x\}|$.

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 that the agent can, in fact, pin down $g_{\mathcal{C}_f}(\theta)$.

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 $x' \in b^*(\theta')$ but $x' \neq g_{\mathcal{C}_f}(\theta')$, then

$$\begin{aligned} |\{g \in B_{-\Theta'} : g(\theta') = x'\}| &= \prod_{\hat{\theta} \in \Theta \backslash (\Theta' \cup \theta')} |b^*(\hat{\theta})| \\ &\leq \prod_{\hat{\theta} \in \Theta \backslash \theta} |b^*(\hat{\theta})| \qquad \text{since } \theta \in \Theta' \\ &\leq K \qquad \text{by (2)} \end{aligned}$$

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

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

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.1.1 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 7 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 $\mathcal{C}_f$. By Lemma 4, either $B_{K,\mathcal{C}} = D(\overline{x})$ or $B_{K,\mathcal{C}} = \{f\}$. If $B_{K,\mathcal{C}} = D(\overline{x})$, then $\mathcal{C}_f$ K-implements f. Otherwise, $B_{K,\mathcal{C}} = \{f\}$ for all K. 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 reach finer beliefs than $D(\overline{x})$, thus triggering Lemma 4). Thus, by part (iv) of Lemma 7, f is trivial. Hence, in all cases, f is implementable.

A.1.2 Proof of Proposition 2 and Corollaries 1 and 2

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

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 satisfy the IR and IC constraints for all K.

If f is nontrivial, apply Lemma 3 to get $B_{K,\mathcal{C}} \subseteq B_{K,\mathcal{C}_f}$. If $K < \overline{K}(\overline{x})$, then $B_{K,\mathcal{C}_f} = D(\overline{x})$ because only an agent of ability $K' \geq \overline{K}(\overline{x})$ 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 7, beliefs $D(\overline{x})$ satisfy the IC and IR constraints, and therefore $\mathcal{C}_f$ K'-implements for all $K' < \overline{K}(\overline{x})$, including K.

If $K > \overline{K}(\overline{x})$, 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 this $\overline{K}(\overline{x})$.

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, $L^*_\theta(\overline{x}') \supseteq L^*_\theta(\overline{x})$ for all θ. It follows that $D(\overline{x}') \subseteq D(\overline{x})$ because $D(\overline{x}')$ has associated correspondence $\hat{b}$ satisfying $\hat{b}(\theta) = X\backslash L^*_\theta(\overline{x}') \subseteq X\backslash L^*_\theta(\overline{x}) = b^*(\theta)$. Clearly, then,

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

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 $L^*_\theta(\overline{x})$ as possible outcomes from reporting θ so that the resulting correspondence b^0 satisfies $b^0(\theta) = X \backslash L^*_\theta(\overline{x})$. 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 8. *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 used in the previous section, the contract

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

either induces belief $b_{K,\mathcal{C}} = D$ or $b_{K,\mathcal{C}} = \{f\}$. Clearly $\mathcal{C}$ achieves K-implementation 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 7), f is trivially strictly implementable: for all θ, $f(\theta) \succ_\theta f(\theta')$ and $f(\theta) \succsim_\theta \overline{x}_\theta$. $\square$

Lemma 9. *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 7). 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, beliefs b^i make response θ' at least as attractive as response θ for type 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 8 and 9 that D^*, the set of all implementable functions, satisfies

$$D^* = \{f : \Theta \to X \mid \forall \theta \in \Theta,\ f(\theta) \in b^{**}(\theta)\} \tag{3}$$

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 8, every $f \in D^*(\overline{x})$ is strictly implementable.

Conversely, by (3), 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.