Chapter 4

Mechanism Design

Honesty is the best policy - when there is money in it.

Mark Twain

In order for a preference aggregator to choose a good outcome, she needs to be provided with the agents' (relevant) preferences. Usually, the only way of learning these preferences is by having the agents report them. Unfortunately, in settings where the agents are self-interested, they will report these preferences truthfully if and only if it is in their best interest to do so. Thus, the preference aggregator has the difficult task of not only choosing good outcomes for the given preferences, but also choosing outcomes in such a way that agents will not have any incentive to misreport their preferences. This is the topic of *mechanism design*, and the resulting outcome selection functions are called *mechanisms*.

This chapter gives an introduction to some basic concepts and results in mechanism design. In Section 4.1, we review basic concepts in mechanism design (although discussions of the game-theoretic justifications for this particular framework, in particular the *revelation principle*, will be postponed to Chapter 7). In Section 4.2, we review the famous and widely-studied *Vickrey-Clarke-Groves* mechanisms and their properties. In Section 4.3, we briefly review some other positive results (mechanisms that achieve particular properties), while in Section 4.4, we briefly review some key impossibility results (combinations of properties that no mechanism can achieve).

4.1 Basic concepts

If all of the agents' preferences were public knowledge, there would be no need for mechanism design—all that would need to be done is solve the outcome optimization problem. Techniques from mechanism design are useful and necessary only in settings in which agents' have *private information* about their preferences. Formally, we say that each agent i has a privately known *type* θ_i that corresponds to that agent's private information, and we denote by Θ_i the space of all of agent i's possible types. In general, it is possible to have private information that has implications for how other agents value outcomes—for example, one agent may privately know that the painting that is being auctioned is a forgery, which would be relevant to other agents that may not know this [Ito *et al.*, 2002, 2003, 2004]. In this dissertation, as is most commonly done in the mechanism design

literature, we will only consider private information about the agent's own preferences (which is the most common type of private information). We model these preferences by saying that each agent i has a utility function $u_i : \Theta_i \times O \to \mathbb{R}$, where $u_i(\theta_i, o)$ gives the agent's utility for outcome o when the agent has type θ_i. The utility function u_i is common knowledge, but it is still impossible for other agents to precisely assess agent i's utility for a given outcome o without knowing agent i's type. For example, in an auction for a single item, an agent's type θ_i could be simply that agent's valuation for the item. Then, the agent's utility for an outcome in which he receives the item will be θ_i (not counting any payments to be made by the agent), and the utility is 0 otherwise. Hence, the utility function is common knowledge, but one still needs to know the agent's type to assess the agent's utility for (some) outcomes.

A *direct-revelation mechanism* asks each agent to report its private information, and chooses an outcome based on this (and potentially some random bits). It will generally be convenient not to consider payments imposed by the mechanism as part of the outcome, so that the mechanism also needs to specify payments to be made by/to agents. Formally:

Definition 16

- *A* deterministic direct-revelation mechanism without payments *consists of an outcome selection function* $o : \Theta_1 \times \ldots \times \Theta_n \to O$.
- *A* randomized direct-revelation mechanism without payments *consists of a distribution selection function* $p : \Theta_1 \times \ldots \times \Theta_n \to \Delta(O)$, *where* $\Delta(O)$ *is the set of probability distributions over* O.
- *A* deterministic direct-revelation mechanism with payments *consists of an outcome selection function* $o : \Theta_1 \times \ldots \times \Theta_n \to O$ *and for each agent* i, *a payment selection function* $\pi_i : \Theta_1 \times \ldots \times \Theta_n \to \mathbb{R}$, *where* $\pi_i(\theta_1, \ldots, \theta_n)$ *gives the payment made by agent* i *when the reported types are* $\theta_1, \ldots, \theta_n$.
- *A* randomized direct-revelation mechanism with payments *consists of a distribution selection function* $p : \Theta_1 \times \ldots \times \Theta_n \to \Delta(O)$, *and for each agent* i, *a payment selection function* $\pi_i : \Theta_1 \times \ldots \times \Theta_n \to \mathbb{R}$.

In some settings, it makes sense to think of an agent's type θ_i as being drawn from a (commonly known) prior distribution over Θ_i. In this case, while each agent still only knows its own type, each agent can use the commonly known prior to make probabilistic assessments of what the others will report.

So, what makes for a good mechanism? Typically, there is an *objective function* that the designer wants to maximize. One common objective is social welfare (the sum of the agents' utilities with respect to their *true*, not reported, types), but there are many others—for example, the designer may wish to maximize revenue (the sum of the agents' payments). However, there are certain constraints on what the designer can do. For example, it would not be reasonable for the designer to specify that a losing bidder in an auction should pay the designer a large sum: if so, the bidder would simply not participate in the auction. We next present constraints, called *participation* or *individual rationality (IR)* constraints, that prevent this. Before we do so, we note that we will assume *quasilinear preferences* when payments are involved.

Definition 17 *An agent i has* quasilinear preferences *if the agent's utility function can be written as $u_i(\theta_i, o) - \pi_i$.*

We are now ready to present the IR constraints.

Definition 18 Individual rationality (IR) *is defined as follows.*

- *A deterministic mechanism is* ex post IR *if for any agent i, and any type vector $(\theta_1, \ldots, \theta_n) \in \Theta_1 \times \ldots \times \Theta_n$, we have $u_i(\theta_i, o(\theta_1, \ldots, \theta_n)) - \pi_i(\theta_1, \ldots, \theta_n) \geq 0$.*

  A randomized mechanism is ex post IR *if for any agent i, and any type vector $(\theta_1, \ldots, \theta_n) \in \Theta_1 \times \ldots \times \Theta_n$, we have $E_{o|\theta_1,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1, .., \theta_n)] \geq 0$.*

- *A deterministic mechanism is* ex interim IR *if for any agent i, and any type $\theta_i \in \Theta_i$, we have $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i}[u_i(\theta_i, o(\theta_1, .., \theta_n)) - \pi_i(\theta_1, .., \theta_n)] \geq 0$.*

  A randomized mechanism is ex interim IR *if for any agent i, and any type $\theta_i \in \Theta_i$, we have $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i} E_{o|\theta_1,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1, .., \theta_n)] \geq 0$.*

The terms involving payments are left out if payments are not possible.

Thus, participating in an *ex post* individually rational mechanism never makes an agent worse off; participating in an *ex interim* individually rational mechanism may make an agent worse off in the end, but not in expectation (assuming that the agent's belief over the other agents' reported types matches the common prior).

Still, as long as these are the only constraints, all that the designer needs to do is solve the outcome optimization problem (perhaps charging the agents' their entire utility as payment, in case revenue maximization is the objective). But we have not yet considered the agents' *incentives*. Agents will only report their preferences truthfully if they have an incentive to do so. We will impose *incentive compatibility (IC)* constraints that ensure that this is indeed the case. Again, there is an *ex post* and an *ex interim* variant; in this context, these variants are usually called *dominant-strategies incentive compatible* and *Bayes-Nash equilibrium (BNE) incentive compatible*, respectively. Given the (potential) difference between true and reported types, we will use the standard notation $\hat{\theta}_i$ to refer to agent i's reported type.

Definition 19 *A mechanism is* dominant-strategies incentive compatible *(or* strategy-proof*) if telling the truth is always optimal, even when the types reported by the other agents are already known. Formally, for any agent i, any type vector $(\theta_1, \ldots, \theta_i, \ldots, \theta_n) \in \Theta_1 \times \ldots \times \Theta_i \times \ldots \times \Theta_n$, and any alternative type report $\hat{\theta}_i \in \Theta_i$, in the case of deterministic mechanisms we require $u_i(\theta_i, o(\theta_1, \ldots, \theta_i, \ldots, \theta_n)) - \pi_i(\theta_1, \ldots, \theta_i, \ldots, \theta_n) \geq u_i(\theta_i, o(\theta_1, \ldots, \hat{\theta}_i, \ldots, \theta_n)) -$ $\pi_i(\theta_1, \ldots, \hat{\theta}_i, \ldots, \theta_n)$. In the case of randomized mechanisms we have $E_{o|\theta_1,..,\theta_i,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1, \ldots, \theta_i, \ldots, \theta_n)] \geq E_{o|\theta_1,..,\hat{\theta}_i,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1, \ldots, \hat{\theta}_i, \ldots, \theta_n)]$.*
The terms involving payments are left out if payments are not possible.

Definition 20 *A mechanism is* Bayes-Nash equilibrium (BNE) incentive compatible *if telling the truth is always optimal to an agent when that agent does not yet know anything about the other*

agents' types, and the other agents are telling the truth. Formally, for any agent i, any type $\theta_i \in \Theta_i$, and any alternative type report $\hat{\theta}_i \in \Theta_i$, in the case of deterministic mechanisms we have $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i}[u_i(\theta_i, o(\theta_1,\ldots,\theta_i,\ldots,\theta_n)) - \pi_i(\theta_1,\ldots,\theta_i,\ldots,\theta_n)] \geq$ $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i}[u_i(\theta_i, o(\theta_1,\ldots,\hat{\theta}_i,\ldots,\theta_n)) - \pi_i(\theta_1,\ldots,\hat{\theta}_i,\ldots,\theta_n)]$*. In the case of randomized mechanisms we have* $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i} E_{o|\theta_1,..,\theta_i,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1,\ldots,\theta_i,\ldots,\theta_n)] \geq$ $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i} E_{o|\theta_1,..,\hat{\theta}_i,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1,\ldots,\hat{\theta}_i,\ldots,\theta_n)]$.

The terms involving payments are left out if payments are not possible.

One may wonder whether it is possible to obtain better outcomes by using a direct-revelation mechanism that is not truthful—perhaps the cost of the resulting strategic misreporting is not as great as the cost of having to honor the incentive compatibility constraints. Or, perhaps we could do even better using a mechanism that is not a direct-revelation mechanism—that is, a mechanism under which agents have other actions to take besides merely reporting their preferences. A famous result called the *revelation principle* [Gibbard, 1973; Green and Laffont, 1977; Myerson, 1979, 1981] shows that, when agents are perfectly strategic (unboundedly rational), the answer to both of these questions is "no": there is no loss in restricting attention to truthful, direct-revelation mechanisms.[1] For now, we do not yet have a definition of strategic behavior in non-truthful or indirect mechanisms, so we will postpone detailed discussion of the revelation principle to Chapter 7 (and we will question the assumption of unbounded rationality in Chapters 8 and 9). However, the intuition behind the revelation principle is simple: suppose we envelop a non-truthful mechanism with an *interface layer*, to which agents input their preferences. Then, the interface layer interacts with the original mechanism on behalf of each agent, *playing strategically in the agent's best interest* based on the reported preferences. (Compare, for example, proxy agents on eBay [eBay UK, 2004].) The resulting mechanism is truthful: an agent has no incentive to misreport to the interface layer, because the layer will play the agent's part in the original mechanism in the agent's best interest. Moreover, the final outcome of the new, truthful mechanism will be the same, because the layer will play strategically optimally—just as the agent would have.

In the next section, we will define the famous Vickrey-Clarke-Groves mechanisms.

4.2 Vickrey-Clarke-Groves mechanisms

The most straightforward direct-revelation mechanism for selling a single item is the *first-price sealed-bid* auction, in which each bidder submits a bid for the item in (say) a sealed envelope, and the highest bidder wins and pays the value that he bid. This is certainly not an incentive-compatible mechanism: in fact, bidding one's true valuation guarantees a utility of 0 (even if the bid wins, the bidder will pay his entire valuation). Rather, to obtain positive utility, a bidder needs to reduce (or *shave*) his bid, ideally to the point where it is only slightly higher than the next highest bid. Another direct-revelation mechanism is the *Vickrey* [Vickrey, 1961] or *second-price sealed-bid* auction, in which the highest bidder still wins, but pays the value of the *second* highest bid. The Vickrey auction is strategy-proof. To see why, imagine a bidder that knows the other bids. This bidder has only two

[1] The result requires that we can use randomized truthful mechanisms. Moreover, if there are multiple strategic equilibria in the original non-truthful mechanism, then we can choose any one of them to be preserved in the truthful mechanism, but not *all* the equilibria are necessarily preserved.

choices: bid higher than the highest other bid, to win and pay the value of that other bid; or bid lower, and do not win the item. The bidder will prefer to do the former if his valuation is higher than the highest other bid, and the latter otherwise. But in fact, bidding truthfully accomplishes exactly this! Hence, bidding truthfully guarantees one the same utility that an omniscient bidder would receive, and therefore the mechanism is strategy-proof.

It turns out that the Vickrey mechanism is a special case of a general mechanism called the *Clarke* mechanism (or Clarke tax) [Clarke, 1971], which can be applied to combinatorial auctions and exchanges, as well as other preference aggregation settings. The Clarke mechanism works as follows. First, choose the optimal outcome based on the bidders' reported preferences; call this outcome o^*. Then, to determine agent i's payment, remove agent i from the preference aggregation problem, and solve this problem again to obtain o^*_{-i}. Agent i will be required to pay $\sum_{j \neq i} u_j(\hat{\theta_j}, o^*_{-i}) - \sum_{j \neq i} u_j(\hat{\theta_j}, o^*)$. Informally, agent i's payment is exactly the amount by which the other agents are worse off due to agent i's presence—the *externality* that i imposes on the other agents. The Clarke mechanism is strategy-proof, for the following reason. Agent i seeks to maximize $u_i(\theta_i, o^*) + \sum_{j \neq i} u_j(\hat{\theta_j}, o^*) - \sum_{j \neq i} u_j(\hat{\theta_j}, o^*_{-i})$. Since o^*_{-i} does not depend on agent i's report, agent i cannot affect the term $\sum_{j \neq i} u_j(\hat{\theta_j}, o^*_{-i})$, so equivalently, agent i seeks to maximize $u_i(\theta_i, o^*) + \sum_{j \neq i} u_j(\hat{\theta_j}, o^*)$. Agent i can only affect this expression by influencing the choice of o^*, and the mechanism will select o^* to maximize $\sum_{j=1}^{n} u_j(\hat{\theta_j}, o^*)$. But then, if the agent reports truthfully, that is, $\hat{\theta_i} = \theta_i$, then the mechanism will choose o^* precisely to maximize $u_i(\theta_i, o^*) + \sum_{j \neq i} u_j(\hat{\theta_j}, o^*)$, thereby maximizing agent i's utility.

The Clarke mechanism is also *ex post* individually rational, *if* 1) the presence of an agent never makes it impossible to choose some outcome that could have been chosen without that agent, and 2) no agent ever has a negative utility for an outcome that would be selected if that agent were not present. Note that if either 1) or 2) does not hold, then the Clarke mechanism may require a payment from an agent that receives a utility of 0 for the chosen outcome, and is therefore not individually rational. Both 1) and 2) will hold in the remainder of this dissertation.

Additionally, the Clarke mechanism is *weak budget balanced*, that is, the sum of the payments from the agents is always nonnegative, *if* the following condition holds: when an agent is removed from the system, the new optimal (welfare-maximizing) outcome is at least as good for the remaining agents as the optimal outcome before the first agent was removed. That is, if an agent leaves, that does not make the other agents worse off in terms of the chosen outcome (not considering payments). This condition does not hold in, for example, task allocation settings: if an agent leaves, the tasks allocated to that agent must be re-allocated to the other agents, who will therefore be worse off. Indeed, in task allocation settings, the agents must be compensated for taking on tasks, so we do not expect weak budget balance. Green and Laffont [1977] show that it is not possible to obtain *strong* budget balance—the sum of the payoffs always being zero—in addition to choosing optimal outcomes and having dominant-strategies incentive compatibility.

Finally, the Clarke mechanism is just one mechanism among the class of *Groves* mechanisms [Groves, 1973]. To introduce this class of mechanisms, we note that in the Clarke mechanism, agent

i's type report $\hat{\theta}_i$ does not affect the terms $\sum_{j \neq i} u_j(\hat{\theta}_j, o^*_{-i})$ in agent i's payment; short of colluding with the other agents, there is nothing that agent i can do about paying these terms. Hence, if we removed these terms from the payment function, the mechanism would still be strategy-proof. Moreover, *any* term that we add to agent i's payment *that does not depend on* $\hat{\theta}_i$ will not compromise strategy-proofness. The class of Groves mechanisms consists precisely of all mechanisms that can be obtained in this manner. Additionally, Groves mechanisms are in fact the *only* mechanisms that are efficient (*i.e.* the mechanism chooses the optimal outcome) and dominant-strategies incentive-compatible, given that there is no restriction on what the agents' types can be [Green and Laffont, 1977] or even only given that agents' type spaces are smoothly connected [Holmström, 1979]. It should be noted that the Clarke mechanism is often referred to as "the" VCG mechanism, and we will follow this convention.

In the next section, I survey some other positive results in mechanism design (without presenting them in full detail).

4.3 Other possibility results

Interestingly, in some settings, there are Groves mechanisms that require a smaller total payment from the agents than the Clarke mechanism, while maintaining individual rationality and never incurring a deficit. The idea here is to *redistribute* some of the Clarke surplus back to the agents. To maintain incentive compatibility, how much is redistributed to an agent cannot depend on that agent's type. Nevertheless, if the other agents' reported types are such that a certain amount of Clarke surplus will be obtained *regardless* of the given agent's report, then we can redistribute a share of that guaranteed surplus (most naturally, $1/n$) to the agent. For example, in a single-item auction, each agent receives $1/n$ of the second-highest bid among the other bids [Cavallo, 2006].

It turns out that if we are willing to use Bayes-Nash incentive compatibility rather than dominant-strategies incentive compatibility, then we can obtain (strong) budget balance, using the dAGVA [d'Aspremont and Gérard-Varet, 1979; Arrow, 1979] mechanism. This mechanism is similar to a Groves mechanism, except that, instead of being paid the sum of other agents' utilities according to their reported types, an agent is paid the *expected* sum of other agent's utilities given only the agent's own report. In addition, payment terms that do not depend on the agent's own report can be set in such a way as to obtain budget balance.

As noted before, maximizing social welfare is not always the objective. Another common objective is to maximize revenue. In the context of auctions, this is often referred to as the problem of designing an "optimal" auction. The Myerson auction [Myerson, 1981] is a general mechanism for maximizing the expected revenue of an auctioneer selling a single item. The Maskin-Riley auction [Maskin and Riley, 1989] generalizes this to the case of multiple units of the same item. Only very limited characterizations of revenue-maximizing combinatorial auctions (with more than one item) are known [Avery and Hendershott, 2000; Armstrong, 2000].

Another positive result exists in the context of voting: if preferences are single-peaked, then choosing the median voter's peak as the winner (as we did in Chapter 2) is a strategy-proof mechanism.

4.4 Impossibility results

In the previous sections, we saw mechanisms that achieve certain sets of desirable properties. In this section, we discuss a few negative results, that state that certain sets of desirable properties cannot be obtained by a single mechanism.

Possibly the best-known impossibility result in mechanism design is the Gibbard-Satterthwaite theorem [Gibbard, 1973; Satterthwaite, 1975]. This result shows a very strong impossibility in very general preference aggregation settings (voting settings). Specifically, it shows that when there are three or more possible outcomes (candidates), two or more agents (voters), and there is no restriction on the preferences that can be submitted (such as single-peakedness), then a (deterministic) mechanism (voting rule) cannot have the following properties simultaneously:

- For every outcome, there exist preference reports by the agents that will make this outcome win.
- The mechanism is non-dictatorial, that is, the rule does not simply always choose a single, fixed voter's most-preferred candidate.
- The mechanism is strategy-proof.

Gibbard [1977] later extended this impossibility result to encompass randomized voting rules as well: a randomized voting rule is strategy-proof only if it is a probability mixture of *unilateral* and *duple* rules. (A rule is unilateral if only one voter affects the outcome, and duple if only two candidates can win.) It is not difficult to see that this result implies the Gibbard-Satterthwaite impossibility result.

As we have seen in the previous section, this impossibility result does not apply in settings where the agents' preferences are more restricted—*e.g.* single-peaked, or quasilinear in settings where payments are possible (in which case VCG can be used). Nevertheless, impossibility results exist in these more restricted settings as well. For example, the Myerson-Satterthwaite impossibility theorem [Myerson and Satterthwaite, 1983] states that even in simple bilateral trade settings with quasilinear utility functions, where we have a single seller with a single item (and a privately held valuation for this item), and a single buyer who may procure the item (and has a privately held valuation for the item), it is impossible to have a mechanism that achieves the following properties simultaneously:

- efficiency (trade takes place if and only if the buyer's valuation for the item is greater than the seller's);
- budget-balance (money may flow between the buyer and the seller, but not from/to other places);
- Bayes-Nash incentive compatibility;
- ex-interim individual rationality.

We will show another, similar impossibility result in Chapter 5.

4.5 Summary

This chapter reviewed basic concepts and results from mechanism design. We first reviewed various types of mechanisms, as well as individual-rationality and incentive-compatibility concepts. We then reviewed Vickrey-Clarke-Groves mechanisms and their properties in detail, and we briefly reviewed some other positive results (that is, mechanisms that achieve certain sets of properties), including Cavallo's redistribution mechanism, the dAGVA mechanism, Myerson and Maskin-Riley auctions, and single-peaked preferences. Finally, we briefly reviewed the Gibbard-Satterthwaite and Myerson-Satterthwaite impossibility results.

Armed with a basic understanding of mechanism design, we are now ready to move on to deeper levels of the hierarchy. The next chapter studies difficulties for classical mechanism design in expressive preference aggregation settings.

Chapter 7

Game-Theoretic Foundations of Mechanism Design

As mentioned in Chapter 4, a result known as the *revelation principle* is often used to justify restricting attention to truthful mechanisms. Informally, it states that, given a mechanism (not necessarily a truthful or even a direct-revelation mechanism) that produces certain outcomes when agents behave strategically, there exists a truthful mechanism that produces the same outcomes. Of course, this informal statement is too unspecific to truly understand its meaning. Which type of truthfulness is obtained—implementation in dominant strategies, Bayes-Nash equilibrium, or something else? More importantly, what exactly does it mean for the agents to "behave strategically"? It turns out that there are really multiple versions of the revelation principle: different types of strategic behavior lead to different types of truthfulness. In this chapter, we will review some basic concepts from game theory, which will provide us with basic definitions of strategic behavior. We will also present two versions of the revelation principle. This will give us a deeper understanding of the motivation for restricting attention to truthful mechanisms, which will be helpful in the next two chapters, where we argue that non-truthful mechanisms need to be considered when agents are computationally bounded.

7.1 Normal-form games

Perhaps the most basic representation of a strategic settings is a game in *normal* or *strategic* form. In such a game, there are n agents (or *players*), and each player i has a set of *strategies* S_i to select from. The players select their strategies simultaneously, and based on this each player i receives a utility $u_i(s_1, \ldots, s_n)$. In the case where $n = 2$ and the number of strategies for each agent is finite, we can represent the game in *(bi)matrix form*. To do so, we label one player the row player, and the other the column player; then, we add a row to the matrix for each row player strategy, and a column for each column player strategy; finally, in each entry of the matrix, we place the players' utilities (starting with the row player's) for the outcome of the game that corresponds to this entry's row and column.

For example, the well-known game of rock-paper-scissors has the following normal-form representation:

	R	P	S
R	0,0	-1,1	1,-1
P	1,-1	0,0	-1,1
S	-1,1	1,-1	0,0

(Note that here, each row and each column is given a label (R, P, S); such labels do not have any strategic importance, so we will sometimes omit them.) Rock-paper-scissors is what is known as a *zero-sum* game, because within each entry of the matrix, the payoffs sum to zero—what one player gains, the other loses. If the payoffs in each entry of the matrix sum to a constant other than zero, the game is effectively still a zero-sum game, because affine transformations of utility do not affect a player's behavior.

7.1.1 Minimax strategies

How should we play rock-paper-scissors (and other zero-sum games)? Let us suppose, pessimistically, that the other player has good insight into how we play. Then, having a deterministic strategy (say, playing "rock" with probability one) is not a good idea, because the other player can play "paper" and win. Instead, it is better to randomize—for example, play each action with probability $1/3$. The set of randomizations ΔS_i over player i's (original) set of strategies in the game is known as the set of player i's *mixed* strategies. (For contrast, we will refer to S_i as the set of *pure* strategies.)

The most conservative way to play a two-player zero-sum game is to assume that the other player is able to predict one's mixed strategy perfectly. Then, one should choose one's mixed strategy to minimize the maximum utility that the other player can obtain (given that that player knows the mixed strategy). Formally, using the common notation $-i$ to denote "the player other than i," player i should choose a strategy from $\arg\min_{\sigma_i \in \Delta(S_i)} \max_{s_{-i} \in S_{-i}} u_{-i}(\sigma_i, s_{-i})$. (When we give a utility function a mixed strategy as an argument, it simply produces the expected utility given that mixed strategy.) This cautious manner of play may appear very favorable to player $-i$ (given that that $-i$ really *does* know player i's strategy). However, in rock-paper-scissors, the minimax strategy is to play each pure strategy with probability $1/3$, and in this case, any action that the opponent takes will result in an expected utility of 0 for both players. So at least in this game, there is no benefit to being able to choose one's strategy based on the opponent's mixed strategy. This is no accident: in fact, the famous Minimax Theorem [von Neumann, 1927] shows that the players' expected utilities will be the same regardless of which player gets to choose last. Formally, we have (if the utilities in each entry sum to 0): $\arg\min_{\sigma_1 \in \Delta(S_1)} \max_{s_2 \in S_2} u_2(\sigma_1, s_2) = -\arg\min_{\sigma_2 \in \Delta(S_2)} \max_{s_1 \in S_1} u_1(\sigma_2, s_1)$. Hence, it is natural to play minimax strategies in two-player zero-sum games.

What if the game is not zero-sum? As will become clear shortly, no perfect generalization of the Minimax Theorem exists; nevertheless, there are still ways of solving these games. One simple, but not always applicable notion is that of *dominance*, which will be discussed in the next section.

7.1.2 Dominance and iterated dominance

Consider the following game, commonly known as the *Prisoner's Dilemma*:

	S	C
S	-1,-1	-3,0
C	0,-3	-2,-2

The story behind the Prisoner's Dilemma game is as follows. Two criminals are arrested in connection with a major crime, but there is only enough evidence to convict them of a minor crime. The criminals are put in separate rooms, and are each given the option of confessing to the major crime (C) or keeping silent (S). If both keep silent, they are convicted of the minor crime and sentenced to one year in prison. If one confesses and the other does not, no charges at all will be filed against the criminal that confesses, and the one that does not is convicted of the major crime and sentenced to 3 years in prison. Finally, if both confess, they are both convicted of the major crime and given a slightly reduced sentence of 2 years in prison.

How should each criminal play? (Note that it is assumed that there is no opportunity for retaliation afterwards, nor do the criminals care about each other's fate—each prisoner's objective is simply to minimize the amount of time that he spends in prison.) If the other criminal confesses, it is better to confess and get -2 rather than -3. But similarly, if the other criminal keeps silent, it is better to confess and get 0 rather than -1. So, confessing is always better, and both criminals will confess—even though this will give each of them a worse outcome than if they had kept silent.[1] We say that confessing is a *dominant strategy*. Formally:

Definition 31 *Player i's strategy $\sigma_i' \in \Delta(S_i)$ is said to be* strictly dominated *by player i's strategy $\sigma_i \in \Delta(S_i)$ if for any vector of strategies $s_{-i} \in S_{-i}$ for the other players, $u_i(\sigma_i, s_{-i}) > u_i(\sigma_i', s{-}i)$. Player i's strategy $\sigma_i' \in \Delta(S_i)$ is said to be* weakly dominated *by player i's strategy $\sigma_i \in \Delta(S_i)$ if*

[1] While prisoners' confessing to a crime may not appear to be such a bad outcome, there are many other real-world strategic situations with roughly the same structure where we clearly would prefer the agents to cooperate with each other and obtain the higher utilities. For example, there are settings where both players would be better off if each invested in a given public good, but if players act selfishly, neither will invest. Perhaps due to the frustrating nature of such outcomes, many suggestions have been made as to why an agent may still choose to act cooperatively. For example, the agents may care about each other's welfare, or bad behavior may cause failed cooperation, or even retaliation, in the future. Such arguments amount to nothing more than saying that the game structure and its utilities are inaccurate (or at least incomplete). Indeed, one should always be careful to model one's setting accurately, but this does not resolve the problem in the many settings that really are modeled accurately by a Prisoner's Dilemma game. A possible exception is the following argument. Suppose a player believes that the other player reasons *exactly* like him, and will therefore always make the same decision. Then, if the former player cooperates, so will the other player; if he does not, neither will the other player. Therefore, the first player should cooperate. This type of reasoning has been called "superrationality" [Hofstadter, 1985], but it quickly leads to difficult questions of causality (does choosing to cooperate "cause" the other player to cooperate?) and free will (is one's decision already pre-ordained given that the other player must do the same?). This is closely related to *Newcomb's paradox* [Nozick, 1969], in which a superintelligent or even omniscient being presents an agent with two boxes, each of which contains some nonnegative amount of money. The agent can choose to take either the contents of the first box only, or the contents of both boxes. The catch is that when filling the boxes, the being predicted whether the agent would take one or both boxes, and if it predicted that the agent would choose only one box, it placed significantly more money in that one box than it otherwise would have placed in both boxes together. Moreover, the being has been absolutely flawless in predicting other, previous agents' choices. It can be argued that the agent should choose only the one box, because then the being presumably would have put much more money in that box; or that the agent should choose both boxes, since the amounts in the boxes are already fixed at this point. In this dissertation, I will not address these issues and simply follow the standard model in which one can make a decision without affecting one's beliefs about what the other players will decide or have decided (which, for most real-world settings, is an accurate model).

for any vector of strategies s_{-i} for the other players, $u_i(\sigma_i, s_{-i}) \geq u_i(\sigma'_i, s_{-i})$, and for at least one vector of strategies s_{-i} for the other players, $u_i(\sigma_i, s_{-i}) > u_i(\sigma'_i, s_{-i})$.

This definition allows the dominating strategy σ_i and the dominated strategy σ'_i to be mixed strategies, although the restriction where these strategies must be pure can also be of interest (especially to avoid assumptions on agents' attitudes towards risk). There are other notions of dominance, such as *very weak* dominance (in which no strict inequality is required, so two strategies can dominate each other), but this dissertation will not study those notions.

In *iterated dominance*, dominated strategies are removed from the game, and no longer have any effect on future dominance relations. For example, consider the following modification of the Prisoner's Dilemma in which the District Attorney severely dislikes the row criminal and would press charges against him even if he were the only one to confess:

	S	C
S	-1,-1	-3,0
C	-2,-3	-2,-2

Now, the dominance argument only works for the column player. However, because (using the dominance argument) it is clear that the column player will not keep silent, that column becomes irrelevant to the row player. Thus the row player effectively faces the following game:

	C
S	-3,0
C	-2,-2

In this remaining game, confessing does once again dominate keeping silent for the row player. Thus, iterated dominance can solve this game completely.

Either strict or weak dominance can be used in the definition of iterated dominance. We note that the process of iterated dominance is never helped by removing a dominated mixed strategy, for the following reason. If σ'_i gives player i a higher utility than σ_i against mixed strategy σ_j for player $j \neq i$ (and strategies $\sigma_{-\{i,j\}}$ for the other players), then for at least one pure strategy s_j that σ_j places positive probability on, σ'_i must perform better than σ_i against s_j (and strategies $\sigma_{-\{i,j\}}$ for the other players). Thus, removing the mixed strategy σ_j does not introduce any new dominances.

7.1.3 Nash equilibrium

Many games cannot be solved using (iterated) dominance. Consider the following game (commonly called "chicken"):

	S	D
S	-2,-2	1,-1
D	-1,1	0,0

The story behind this game is the following: to test who has the strongest nerves, two drivers drive straight at each other, and at the last moment each driver must decide whether to continue straight (S) or dodge the other car by turning (say) right (D). The preferred outcome is to "win" by going straight when the other dodges, but if both drivers continue straight, they collide and both suffer severely.

This game has no dominated strategies. In fact, the matrix has multiple strategically stable entries: if one player goes straight, and the other dodges, then neither player has an incentive to change strategies (the player going straight is winning, and the player dodging does not want to go straight and collide). This leads to the definition of a *Nash equilibrium*:

Definition 32 *Given a normal-form game, a* Nash equilibrium *is vector of mixed strategies $\sigma_1, \ldots, \sigma_n$ such that no agent has an incentive to deviate from its mixed strategy given that the others do not deviate. That is, for any i and any alternative mixed strategy σ_i', we have $u_i(\sigma_1, \ldots, \sigma_i, \ldots, \sigma_n) \geq u_i(\sigma_1, \ldots, \sigma_i', \ldots, \sigma_n)$.*

Indeed, (S, D) and (D, S) are pure-strategy Nash equilibria of "chicken." There is another Nash equilibrium where both players play each pure strategy with probability 0.5. Every finite game has at least one Nash equilibrium if we allow for mixed strategies [Nash, 1950].

7.2 Bayesian games

The normal-form representation of games assumes that players' utilities for outcomes of the game are common knowledge. Hence, they cannot directly capture settings in which the players' have *private information* about their utilities, as they would, for example, in an auction. Such settings can be modeled using *Bayesian games*.

In a Bayesian game, each player first receives privately held preference information (the player's *type*) from a distribution, which determines the utility that that player receives for every outcome of (that is, vector of actions played in) the game. After receiving this type, the player plays an action based on it.[2]

Definition 33 *A* Bayesian game *is given by a set of players $\{1, 2, \ldots, n\}$; and, for each player i, a set of actions A_i, a type space Θ_i with a probability distribution p_i over it, and a utility function $u_i : \Theta_i \times A_1 \times \ldots \times A_n \to \mathbb{R}$ (where $u_i(\theta_i, a_1, \ldots, a_n)$ denotes player i's utility when i's type is θ_i and each player j plays action a_j). A* pure strategy *in a Bayesian game is a mapping from types to actions, $s_i : \Theta_i \to A_i$, where $s_i(\theta_i)$ denotes the action that player i plays for type θ_i.*

As an example, consider an unusual first-price sealed-bid auction with two bidders, in which the bidders can only bid 1 or 2. If the bids are tied, then the winner is chosen randomly. Each bidder draws a valuation from $\Theta_1 = \Theta_2 = \{2, 2.5\}$ uniformly at random. We can represent the utility function of player 1 (the row player) as follows:

[2] In general, a player can also receive a signal about the other players' preferences, but we will not concern ourselves with that in this dissertation.

	bid 1	bid 2
bid 1	.5	0
bid 2	0	0

Row player utilities when $\theta_1 = 2$.

	bid 1	bid 2
bid 1	.75	0
bid 2	.5	.25

Row player utilities when $\theta_1 = 2.5$.

The utility function for the column player is similar.

Any vector of pure strategies in a Bayesian game defines an (expected) utility for each player, and therefore we can simply translate a Bayesian game into a normal-form game. For example, the auction game above gives (letting x, y denote the strategy of bidding x when one's type is 2, and y when one's type is 3):

	1,1	1,2	2,1	2,2
1,1	.625, .625	.3125, .5	.3125, .375	0, .25
1,2	.5, .3125	.3125, .3125	.3125, .1875	.125, .1875
2,1	.375, .3125	.1875, .3125	.1875, .1875	0, .1875
2,2	.25, 0	.1875, .125	.1875, 0	.125, .125

Using this transformation, we can take any solution concept that we have defined for normal-form games (such as dominance or Nash equilibrium), and apply it to Bayesian games. For example, in the game above, the strategy 2,1 is strictly dominated by 1,2. The strategy 2,2 is weakly dominated by 1,2. After removing 2,2 for both players, 1,1 weakly dominates every other strategy, so iterated weak dominance can solve this game entirely, leaving only 1,1 for each player. Both players playing 2,2 is nevertheless a Nash equilibrium so is both players playing 1,2; and both players playing 1,1. There are no mixed-strategy equilibria.

One remark that should be made is that the normal-form representation of the Bayesian game is exponentially larger than the original representation, because each player i has $|A_i|^{|\Theta_i|}$ distinct pure strategies. For the purpose of defining solution concepts and other conceptual purposes, this causes no problem. But, later, when we will be interested in computing Bayesian games' solutions, it will not be sufficient to simply apply this transformation and solve the normal form, since this will require exponential time (and space).

So, one can define solution concepts for Bayesian games by applying normal-form solution concepts to the normal-form representation of a Bayesian game. In spite of the simplicity of this approach, the typical approach in mechanism design is nevertheless to define the solution concepts directly, as is done below. For simplicity of notation, in the remainder of this chapter, I discuss pure strategies only; the generalizations to mixed strategies (where agents choose a distribution over actions based on their types) are straightforward.

First, let us consider a direct definition of dominance that is typically used in mechanism design:

Definition 34 *Given a Bayesian game, the vector of strategies* $(s_1, \ldots, s_n)$ *is a* dominant-strategy equilibrium *if for every agent* i*, for every type* $\theta_i \in \Theta_i$*, every alternative action* $a_i \in A_i$*, and every action vector* $a_{-i} \in A_{-i}$ *of the other agents, we have* $u_i(\theta_i, s_i(\theta_i), a_{-i}) \geq u_i(\theta_i, a_i, a_{-i})$.

There are a few differences between this definition and using the normal-form representation definition of dominance given above. First, this definition only applies to games where each agent has a strategy that dominates all others, *i.e.* dominance can solve the game entirely (without iteration). Second, none of the inequalities are required to be strict—this is *very weak* dominance. A third, subtle, minor difference is that in this definition the strategy is supposed to give an optimal action *for every type of the agent*, against any opponent actions. The definition that appeals to the normal-form representation only requires that the strategy maximizes the *total expected utility over the agent's types*, against any opponent actions. The normal-form definition still requires that the strategy chooses the optimal action for any type with positive probability; the only difference is that the normal-form definition does not require optimal actions to be chosen on types that have probability zero. For games with finitely many types, this is an insignificant difference, since it does not make sense to even bother defining a type that occurs with zero probability. Under continuous type spaces, the difference is a little more significant since the normal-form definition may choose to play in a bizarre manner on a set of types with measure zero. Since we will be mainly concerned with finite type spaces, the difference between the definitions is immaterial.

Now we will consider *Bayes-Nash equilibrium*, under which agents strategies are optimal only given the other agents' strategies, and given that one does not know the other agents' types.

Definition 35 *The vector of strategies* $(s_1, \ldots, s_n)$ *is a* Bayes-Nash equilibrium *if for every agent* i*, for every type* $\theta_i \in \Theta_i$*, and every alternative action* $a_i \in A_i$*, we have* $E_{\theta_{-i}}[u_i(\theta_i, s_i(\theta_i), s_{-i}(\theta_{-i}))] \geq E_{\theta_{-i}}[u_i(\theta_i, a_i, s_{-i}(\theta_{-i}))]$.

This definition is identical to the one where we simply apply Nash equilibrium to the normal form of the Bayesian game—with the exception that agents can no longer behave arbitrarily for types that have zero probability.

Now that we have some methods for predicting strategic behavior in arbitrary games, we can return to mechanism design and begin to assess the quality of mechanisms that are not truthful, direct-revelation mechanisms. In the next section, we will use this ability to prove two variants of the revelation principle, showing that *if* agents play according to the solution concepts defined here, then there is no reason not to use a truthful, direct-revelation mechanism.

7.3 Revelation principle

To prove the revelation principle, we first need to assess what outcomes will be produced by a mechanism that is not a truthful, direct-revelation mechanism, based on the solution concepts for Bayesian games given above. Such a mechanism can be represented by a set of actions A_i for each agent i, and an outcome selection function $o : A_1 \times \ldots \times A_n \rightarrow O$. (To minimize notational overhead, payments should be considered part of the outcome here. Also, the outcome function may in general produce distributions over outcomes; everything below can be extended to allow for this as well simply by replacing O with $\Delta(O)$.)

We first define when a mechanism *implements* a given *social choice rule*:

Definition 36 *A* social choice rule *is a function* $f : \Theta_1 \times \ldots \times \Theta_n \rightarrow O$*. A mechanism* o implements *rule* f in dominant strategies *if there is a dominant strategy equilibrium* $(s_1, \ldots, s_n)$ *such that for*

all $(\theta_1, \ldots, \theta_n) \in \Theta_1 \times \ldots \times \Theta_n$, $o(s_1(\theta_1), \ldots, s_n(\theta_n)) = f(\theta_1, \ldots, \theta_n)$. *Similarly, a mechanism* o *implements rule* f in Bayes-Nash equilibrium *if there is a Bayes-Nash equilibrium* $(s_1, \ldots, s_n)$ *such that for all* $(\theta_1, \ldots, \theta_n) \in \Theta_1 \times \ldots \times \Theta_n$, $o(s_1(\theta_1), \ldots, s_n(\theta_n)) = f(\theta_1, \ldots, \theta_n)$.

One should note that a game may have multiple equilibria, and may therefore implement multiple social choice rules. For example, consider the two-type first-price auction example in the previous section: two of its equilibria always allocate the item at random, but the third allocates the item to the bidder with the higher valuation if the valuations are not equal. If there are multiple equilibria, then we will assume that we can choose our favorite equilibrium. This strengthens the power of indirect/non-truthful mechanisms, and therefore strengthens the revelation principle result below. (It should be remarked that truthful direct-revelation mechanisms may have multiple equilibria as well; however, one may argue that the truth-telling equilibrium is "focal", *i.e.* the most natural one.)

We are now ready to review two known variants of the revelation principle, corresponding to dominant-strategies equilibrium and Bayes-Nash equilibrium. Before doing so, recall the simple intuition behind the revelation principle: the new, truthful direct-revelation mechanism that we construct requests the agents' types, and then plays "on their behalf" in the old mechanism (according to the equilibrium of that mechanism) to produce the outcome. There is no reason for an agent to misreport his type, since this will only result in the new mechanism playing the part of that agent suboptimally in the old mechanism.

Revelation Principle, version 1 *Suppose there is an (indirect/non-truthful) mechanism that implements social choice rule* f *in dominant strategies. Then there exists a dominant-strategies incentive-compatible direct-revelation mechanism with outcome selection function* o *that also implements* f *in dominant strategies (using the truth-telling equilibrium).*

Proof: We show how to transform the given mechanism that implements f into a truthful direct-revelation mechanism that implements f. For each i, let $s_i^{old} : \Theta_i \rightarrow A_i^{old}$ be the strategy played by agent i in the equilibrium that implements f in the given mechanism, and let o^{old} be the given game's outcome selection function, so that $o^{old}(s_1^{old}(\theta_1), \ldots, s_n^{old}(\theta_n)) = f(\theta_1, \ldots, \theta_n)$, and the s_i^{old} constitute a dominant strategies equilibrium. Then let our new mechanism have the outcome function o given by $o(\theta_1, \ldots, \theta_n) = o^{old}(s_1^{old}(\theta_1), \ldots, s_n^{old}(\theta_n)) = f(\theta_1, \ldots, \theta_n)$. All we need to show is that truthtelling is a dominant strategies equilibrium. To show this, we observe that for any i and $\theta_i \in \Theta_i$, for any alternative type $\hat{\theta}_i \in \Theta_i$, and for any $\theta_{-i} \in \Theta_{-i}$, $u_i(\theta_i, o(\theta_i, \theta_{-i})) = u_i(\theta_i, o^{old}(s_i^{old}(\theta_i), s_{-i}^{old}(\theta_{-i}))) \geq u_i(\theta_i, o^{old}(s_i^{old}(\hat{\theta}_i), s_{-i}^{old}(\theta_{-i}))) = u_i(\theta_i, o(\hat{\theta}_i, \theta_{-i}))$, where the inequality derives from the fact that the s_i^{old} constitute a dominant strategies equilibrium in the original mechanism. ■

Revelation Principle, version 2 *Suppose there is an (indirect/non-truthful) mechanism that implements social choice rule* f *in Bayes-Nash equilibrium. Then there exists a Bayes-Nash equilibrium incentive-compatible direct-revelation mechanism with outcome selection function* o *that also implements* f *in Bayes-Nash equilibrium (using the truth-telling equilibrium).*

Proof: We show how to transform the given mechanism that implements f into a truthful direct-revelation mechanism that implements f. For each i, let $s_i^{old} : \Theta_i \rightarrow A_i^{old}$ be the strategy played

by agent i in the equilibrium that implements f in the given mechanism, and let o^{old} be the given game's outcome selection function, so that $o^{old}(s_1^{old}(\theta_1), \ldots, s_n^{old}(\theta_n)) = f(\theta_1, \ldots, \theta_n)$, and the s_i^{old} constitute a Bayes-Nash equilibrium. Then let our new mechanism have the outcome function o given by $o(\theta_1, \ldots, \theta_n) = o^{old}(s_1^{old}(\theta_1), \ldots, s_n^{old}(\theta_n)) = f(\theta_1, \ldots, \theta_n)$. All we need to show is that truthtelling is a Bayes-Nash equilibrium. To show this, we observe that for any i and $\theta_i \in \Theta_i$, for any alternative type $\hat{\theta}_i \in \Theta_i$, $E_{\theta_{-i}}[u_i(\theta_i, o(\theta_i, \theta_{-i}))] = E_{\theta_{-i}}[u_i(\theta_i, o^{old}(s_i^{old}(\theta_i), s_{-i}^{old}(\theta_{-i})))] \geq E_{\theta_{-i}}[u_i(\theta_i, o^{old}(s_i^{old}(\hat{\theta}_i), s_{-i}^{old}(\theta_{-i})))] = E_{\theta_{-i}}[u_i(\theta_i, o(\hat{\theta}_i, \theta_{-i}))]$, where the inequality derives from the fact that the s_i^{old} constitute a Bayes-Nash equilibrium in the original game. ■

We have assumed that the strategies in the equilibrium of the original mechanism are pure; the result can be extended to the setting where they are mixed. In this case, though, the resulting truthful mechanism may become randomized, even if the original mechanism was not.

7.4 Summary

In this chapter we reviewed basic concepts from game theory. We reviewed basic solution concepts for normal-form games, including minimax strategies, dominance and iterated dominance, and Nash equilibrium. We then showed how to extend these solution concepts to Bayesian games. Armed with these concepts, we finally presented the (known) proofs of two variants of the *revelation principle*, which (informally stated) show that if agents act strategically (according to these solution concepts), then there is no reason not to use a truthful, direct-revelation mechanism.

Unfortunately, as we will see in the next chapters, the assumption that agents will behave in a strategically optimal way is often untenable in mechanisms for expressive preference aggregation. This is in part due to the fact that the agents' strategy spaces become too large to search exhaustively. Of course, exhaustive search is not necessarily required to behave in a strategically optimal way—perhaps there are efficient algorithms that home in on the optimal strategies quickly. In Chapter 8 we show that for some settings, this is unlikely to be the case, because even the problem of finding a best response to given strategies by the other players is computationally hard (NP-complete or harder). Additionally, intuitively, the problem of computing a best response is much easier than that of acting optimally when the other agents' actions are not yet known, and must be reasoned about first. In Chapter 9 we show that indeed, standard solution concepts such as (iterated) dominance and Nash equilibrium can be hard to compute (even when the strategy spaces are much more manageable in size).