Implementation with Boundedly Rational Agents

Kemal Yildiz*

August 31, 2017

A common underlying assumption in mechanism design literature is the rationality of the agents. That is, agents take actions as to maximize their preferences. However, ample evidence in marketing, psychology and behavioral economics shows that people's choices are not always consistent with the maximization of a preference relation. This inconsistency is of concern to principals, policymakers, institutional designers who aim to implement their goals by designing mechanisms for "real" agents who may not be rational, but may follow simple heuristics, rules of thumb in taking actions. The relevance and importance of taking bounded rationality into account when designing mechanisms is attracting greater attention especially since Thaler and Sunstein's famous book Nudge [9]. However, only recently have we been presented with formal mechanism design models that incorporate bounded rationality. Here, we first present a brief overview and classification of these models. Then, we formally present a class of mechanisms, introduced by Koray and Yildiz [6], that incorporates bounded rationality into the implementation problem.

First, to give an overview of the implementation problem, we start with a social objective -mostly represented by a social choice rule- that is assumed to have been unanimously agreed upon in the society. Moreover, the objective to be implemented relates the social desirability of an outcome to the current societal preferences and other relevant parameters in a precise manner. It is exactly this ambitious aim, which gives rise to the main problem concerning implementation, namely the designer lacks the ability to observe the individuals' actual preferences. To deal with this problem, Hurwicz [4] introduced the notion of a mechanism. In the past three decades, many successful results have been obtained in identifying the set of social choice rules implementable according to widely used game theoretic equilibrium notions. In all these studies a common assumption is the rationality of the agents.

*Bilkent University, Department of Economics, kemal.yildiz@bilkent.edu.tr

On the other hand, a recent and growing body of research, boundedly rational choice theory, seeks to explain documented choice behavior that a rational agent does not exhibit. A boundedly rational choice procedure, which an agent follows to make a choice, might contain a component that can be interpreted as the welfare preference of the agent, as well as components like an alternative that serves as a *status quo*, a *list* that provides an ordering to encounter the outcomes, a second relation used to shortlist some of the outcomes for a final choice, or an *attention filter* that provides the set of outcomes that grabs the attention of the decision maker at a first glance.

In mechanism design literature, bounded rationality is typically incorporated as to create an additional challenge, in which the designed mechanisms should implement the desired goals even if agents fail to behave rationally. Put differently, the designed mechanism should be robust to bounded rationality. However, as recently exemplified by Glazer and Rubinstein [3], the designer can nudge the agents as to benefit from the bounded rationality of the agents and implement the goals that he cannot implement with rational agents via using classical tools such as costly screening or monetary transfers.

We observe that the bulk of research in mechanism design with boundedly rational agents can be classified as either following, what we call, the *robustness approach* or the *nudging approach*. First let us briefly describe these two related, but different approaches.

Robustness approach: In this approach the designer should design a robust mechanism that would implement his goals even if the agents fail to behave rationally. Hence, the bounded rationality of the agents creates an additional challenge for the designer.

Nudging approach: In the nudging approach the mechanism designer implements his goals by taking the advantage of agent's bounded rationality, and implements the goals that he may not implement with rational agents. The nudging approach facilitates the implementation problem, since the designer has the additional power to nudge the agents by designing the external conditions such as the default action, the shortlisting rationale or the list that guides the agents in choosing their actions. Hence, nudging approach creates the possibility of implementation even if the designer does not have the traditional tools for implementation.

To best of our knowledge the first model that incorporates bounded rationality into mechanism design is offered by Eliaz [2], who formulated an implementation model that allows *faulty* agents. In classical implementation framework preferences have two roles. First, given a social choice rule, a preference realization determines the set of acceptable outcomes. Second, given a mechanism,

we assume that each agent takes an action so to maximize his preference relation. In *fault tolerant implementation* model of Eliaz [2], agents are endowed with preference relations, which are used to determine the set of acceptable outcomes as usual. However, not all the agents take actions so to maximize their preferences. The agents who do not act according to their preferences are called faulty. In this setting, the aim of the principal is to implement the set of acceptable outcomes regardless of the number and the identity of these faulty agents. This model perfectly exemplifies what we name as the robustness approach since the designer should design a mechanism for implementation that would tolerate arbitrary behavior of faulty agents.

In a rather recent paper, de Clippel [1] provides a general framework, called *behavioral implementation*. Given a class of boundedly rational choice procedures, it is assumed that each agent is endowed with a choice procedure that belongs to this class. This leads to two departures from the classical setup. First, a social choice rule determines a set of acceptable outcomes for each choice function profile. Second, given a mechanism, each agent chooses an action according to his realized choice function. As in Eliaz [2], the task of the principal is to implement the social goals given that each agent follows a procedure that belongs to the given class. Hence, this model is also more in line with the robustness approach.

In contrast to these two models, Glazer and Rubinstein [3] propose a model in which the principal does not have the traditional tools for implementation, such as costly screening or monetary incentives, yet he can still implement his goals by nudging, that is by taking the advantage of agent's bounded rationality. Glazer and Rubinstein present a persuasion situation as a leader-follower relation. In this model, the persuasion rule and its frame is determined by the leader. The actions that the follower shows attention are restricted depending on how the persuasion rule is framed by the leader.

As complementary to their main model, Koray and Yildiz [6] formulate a general strategic framework that incorporates bounded rationality into the implementation problem in line with the nudging approach. In this framework, the design object is called a *deviation-constrained mechanism* (dc-mechanism). In a classical mechanism, an agent can deviate from a joint strategy by choosing any strategy from his strategy set, as he is independent of the joint strategy from which the deviation is to be made. On this front, a dc-mechanism is different from a classical mechanism. Namely, deviation strategies of each agent are constrained depending on the joint strategy from which the deviation is to be made. Next, we introduce some necessary notation, and then formally present implementation via dc-mechanisms.

We use A to denote a nonempty, finite set of alternatives, and N a nonempty finite set of agents. For given A and N, for each $i \in N$, $\succ_i$ denotes the **preference relation** of agent i – a complete, transitive, antisymmetric binary relation – on A. For each distinct pair $a, b \in A$, $a \succ_i b$ denotes that "i prefers a to b". A **preference profile** $\succ = (\succ_1, \ldots, \succ_n)$. The collection of all preference profiles is denoted by $\mathcal{P}$. A **social choice rule** (SCR) F maps each preference profile into a nonempty subset of A. An SCR F is **unanimous** if for each $a \in A$ we have $F(\succ) = \{a\}$ whenever every agent in the society prefers a to all other alternatives.

A **dc-mechanism** is a triple $(M, \mathcal{D}, g)$. As in the classical mechanism, the **joint strategy set** $M = \Pi_{i \in N} M_i$, where M_i stands for the **strategy set** of agent i. The **outcome function** g maps every joint strategy to an alternative, i.e. $g : M \to A$. For each agent $i \in N$, the **constraint function** $\mathcal{D}_i$ maps each joint strategy m to a subset of M_i, i.e. $\mathcal{D}_i : M \to M_i$. In a dc-mechanism, if an agent i would deviate from strategy m, he is constrained to choosing his strategy from $\mathcal{D}_i(m)$. Given a preference profile $\succ$, a joint strategy m is an **equilibrium** of the dc-mechanism $(M, \mathcal{D}, g)$ at $\succ$ if for each $i \in N$ and $m_i' \in \mathcal{D}_i(m)$, $g(m) \succ_i g(m_i', m_{-i})$. We denote the equilibria of $(M, \mathcal{D}, g)$ at $\succ$ by $\mathbf{E}(M, \mathcal{D}, g, \succ)$.

Definition 1 *A SCR F is implementable via a dc-mechanism if there exists a dc-mechanism $(M, \mathcal{D}, g)$ such that for each $\succ \in \mathcal{P}$, $F(\succ) = g(\mathbf{E}(M, \mathcal{D}, g, \succ))$*

In classical implementation via mechanisms, Maskin [8] shows that every Nash-implementable SCR is monotonic, and monotonicity combined with *no veto power* condition is sufficient for Nash implementability in the presence of at least three agents. It follows from the findings of Koray and Yildiz [6] that Maskin monotonicity is both necessary and sufficient for a unanimous SCR to be implementable via a dc-mechanism in the presence of at least three alternatives. Before proceeding to the definition of Maskin monotonicity, let us introduce some useful notation. The **lower contour set of $\succ_i$ with respect to** $a \in A$, denoted by $\mathrm{L}(\succ_i, a)$, is the set of alternatives to which a is preferred by agent i, i.e. $\mathrm{L}(\succ_i, a) = \{b \in A : a \succ_i b\}$. An SCR F is **Maskin monotonic** if for each $\succ^1, \succ^2 \in \mathcal{P}$ and $a \in F(\succ^1)$, we have $a \in F(\succ^2)$ whenever for every $i \in N$, $\mathrm{L}(\succ_i^1, a) \subseteq \mathrm{L}(\succ_i^2, a)$.

Proposition 1 *In the presence of at least three agents, a unanimous SCR F is implementable via a dc-mechanism if and only if F is Maskin monotonic.*

Proof. Directly follows from Proposition 1 and Proposition 6 of Koray and Yildiz [6]. ■

Specific forms of deviation constraints have been considered previously in the mechanism design literature. For example, Hurwicz, Maskin, and Postlewaite [5] assume that an agent can claim to have any subset of his true endowment but can not claim a larger set of endowments in a pure exchange economy setting. In the model of Glazer and Rubinstein [3] that is discussed earlier, the persuasion rule and its frame is determined by the leader, and, as in a dc-mechanism, the strategies that the follower can choose are restricted depending on how the persuasion rule is framed by the leader. Implementation with dc-mechanisms differs from the *behavioral implementation* model of de Clippel [1] in two main directions. First, de Clippel assumes that an SCR aggregates a choice function profile. In contrast, here an SCR aggregates a preference profile as usual. Secondly, in both models, given a mechanism, each agent can be boundedly rational in choosing an action. However, in dc-mechanisms it is assumed that the attention filter of the agents are also subject to design. Thus, one can implement social choice rules that can be implemented neither in the classical setup, nor in the setup of de Clippel.

From the bounded rationality perspective, implementation via dc-mechanisms proposes a general framework in which agents may exhibit *limited attention* –in the vein of Masatlioglu et al [7]– to their available strategies. In this framework the designer can additionally design the actions that each agent shows *attention* whenever he considers a deviation from a given joint action. Assuming that the designer can shape the actions that grab the attention of each agent in any arbitrary way can be too general and unrealistic. However, Proposition 1 shows that even this unrealistic freedom of designing the deviation constraints does not bring much in terms of implementability. This is because deviation constraints should be both stringent enough to keep a particular strategy when it is an equilibrium at a preference profile, and also permissive enough to eliminate the same strategy when it is not an equilibrium at another preference profile.

As for the future studies, rather specific forms of dc-mechanisms can be formulated in which agents follow different boundedly rational choice procedures. In these models, the designer can be endowed with the additional power of *nudging* or completely designing the primitives such as a *status quo*, a *list*, or an *attention filter* that are relevant for the strategic choice of the agents. It would be an interesting and insightful theoretical exercise to see which SCRs are implementable via these models.

References

[1] Geoffroy de Clippel, *Behavioral implementation*, The American Economic Review **104** (2014), no. 10, 2975–3002. 3, 5

[2] Kfir Eliaz, *Fault tolerant implementation*, The Review of Economic Studies **69** (2002), no. 3, 589–610. 2, 3

[3] Jacob Glazer and Ariel Rubinstein, *A model of persuasion with a boundedly rational agent*, forthcoming, Journal of Political Economy. 2, 3, 5

[4] Leonid Hurwicz, *Decision and organization,*, ch. On Informationally Decentralized Systems, Amsterdam: North Holland, 1972. 1

[5] Leonid Hurwicz, Eric Maskin, and Andrew Postlewaite, *Feasible nash implementation of social choice rules when the designer does not know endowments or production sets*, The Economics of Informational Decentralization: Complexity, Efficiency, and Stability, Springer, 1995, pp. 367–433. 5

[6] Semih. Koray and Kemal Yildiz, *Implementation via right structures*, Bilkent University, mimeo. 1, 3, 4

[7] Yusufcan Masatlioglu, Daisuke Nakajima, and Erkut Y Ozbay, *Revealed attention*, The American Economic Review **102** (2012), no. 5, 2183–2205. 5

[8] Eric Maskin, *The theory of implementation in Nash equilibrium: A survey*, Social Goals and Social Organization (L. Hurwicz, D. Scheidler, and H. Sonnenschein, eds.), Cambridge University Press, 1985. 4

[9] R. H. Thaler and C. R. Sunstein, *Nudge: Improving decisions about health, wealth, and happiness*, Yale University Press, 2008. 1