Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2010/47

Pareto-Optimal Matching Allocation Mechanisms for Boundedly Rational Agents

Sophie Bade

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2010/47

Pareto-Optimal Matching Allocation Mechanisms for Boundedly Rational Agents

Sophie Bade

December 2010

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Pareto-Optimal Matching Allocation Mechanisms for Boundedly Rational Agents

Sophie Bade*†

December 5, 2010

Abstract

This article is concerned with the welfare properties of trade when the behavior of agents cannot be rationalized by preferences. I investigate this question in an environment of matching allocation problems. There are two reasons for doing so: firstly, the finiteness of such problems entails that the domain of the agents' choice behavior does not need to be restricted in any which way to obtain results on the welfare properties of trade. Secondly, some matching allocation mechanisms have been designed for non-market environments in which we would typically expect boundedly rational behavior. I find qualified support for the statements that all outcomes of trade are Pareto-optimal and all Pareto optima are reachable through trade. Contrary to the standard case, different trading mechanisms lead to different outcome sets when the agents' behavior is not rationalizable. These results remain valid when restricting attention to "minimally irrational" behavior.

Keywords: Fundamental Theorems of Welfare, House Allocation Problems, Bounded Rationality, Multiple Rationales. *JEL Classification Numbers*: C78, D03, D60.

*Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. bade@coll.mpg.de

†I would like to thank Christoph Engel, Olga Gorelkina, Martin Hellwig, Aniol Llorente-Saguer, and the seminar audience at Cergy-Pontoise for their comments. I would like to thank David Bielen for thorough research assistance.

1 Introduction

Trade leads to Pareto-optimal outcomes and any Pareto-optimal outcome can be reached via trade. The First and Second Fundamental Theorems of Welfare Economics state sets of conditions under which these two statements hold true. In this article, I investigate the question whether and how the assumption of individual rationality is needed to obtain these results. In other words: I ask whether the two introductory statements on the Pareto optimality of trade can hold true without the assumption that the agents' behavior is rationalizable, in the sense that their choice functions can be derived from the maximization of some (transitive and complete) preference relation.[1]

I treat this question within a framework of matching allocation problems, in which some indivisible objects need to be assigned to some agents. In such environments, trade can be identified with trading-cycles-mechanisms (more on that in Section 4). Two main reasons underlie my choice of this environment: first of all, such matching allocation problems are finite. This entails that *no* assumptions on preferences are needed to show that trade leads to Pareto-optimal outcomes and that any Pareto-optimal outcome can be reached via trade in the standard case in which all agents' behavior is rationalizable (see Abdlukadiroglu and Sonmez [1] and Bade [6]). The same holds true for the welfare theorems proposed in the present paper; they hold for *all* possible choice functions. This is important since assumptions such as local non-satiation or convex upper contour sets are − if anything − harder to interpret and justify in an environment of boundedly rational behavior.[2]

Secondly, some of the non-market environments for which economists have designed matching mechanisms can serve as prime examples of cases in which to expect non-rationalizable behavior. Take kidney allocation problems as an example. One difficulty with the implementation of mechanisms that match donors to recipients is that doctors are reluctant to state complete and transitive preferences over kidneys. However, the same doctors do not seem to have any problem choosing the "best" kidney for a particular patient from a given set.[3] This apparent

[1] The terms "rationalize" or "rationalizable" carry two different meanings in economic theory: in game theory, a strategy profile is considered "rationalizable" if it survives the sequential elimination of dominated strategies. In decision theory, a choice correspondence is considered "rationalizable" if there exists some transitive and complete preference relation $\succsim$, such that the choice correspondence maps any consumption set S to the set of $\succsim$-maximal elements in that set S. In this article, I only use the terms in the second sense.

[2] Some of the eminent studies of the equilibria of competitive markets with boundedly rational agents impose such conditions on the behavior of agents to obtain results; see Fon and Otani [13], Gale and Mas Colell [14], and Mandler [18].

[3] These statements reflect a private conversation with Utku Unver, who was involved in the design and practical implementation of several kidney exchange mechanisms.

contradiction could arise from doctors having only limited resources to test for quality of kidneys.[4] If doctors are aware that their decision procedures can lead to non-rationalizable choices, it is only reasonable for them to refuse to state complete rankings over kidneys.

Alternatively, consider the allocation of elementary school slots to students. Consider the case in which decisions over schooling are not taken by a single agent, but arise out of the interaction of some competing interests. One could think of a situation in which a mother whittles down the available options to be presented to the father, who then chooses among them. Choices that arise from such strategic interplay are generally not rationalizable.[5] So, given that the assumption of preference maximization appears strong in some environments in which trading mechanisms have been implemented, we should ask whether the results on the Pareto optimality of these mechanisms extend to environments with boundedly rational agents.

Some major hurdles need to be cleared before I can go on to state and prove the main results of the present article: the first one concerns the fact that the notion of Pareto optimality builds on the notion of individual preferences. This is problematic, since the very purpose of the present paper is to cover behavior that cannot be explained by preference maximization. In Section 2 I define Pareto optimality in terms of two alternative notions of "individual preference". I say that an agent solidly prefers some object x to another object y if he never chooses y when x is also available. Conversely, an agent lightly prefers x to y if he chooses x out of at least one set that also contains y. Each of these two notions of preference yields a different notion of Pareto optimality. For any matching allocation problem, the resulting two sets of Pareto optima are nested.

The next hurdle is cleared in Section 4, which is concerned with a notion of "free trade" that applies to an environment without money. In tune with the literature on housing problems, I identify "trade" with mechanisms that assign property rights over all objects at the outset and then let agents freely exchange houses in "trading cycles". One last hurdle remains: some

[4]The following story might explain the contrast between the hesitance to state preferences and the readiness to choose. Consider the task to find the "best" kidney for patient x out of a set of ten kidneys. Financial constraints might force doctors to use some preliminary quick and cheap tests, to limit the set of kidneys on which they run some more detailed and expensive tests. Call the kidney chosen according to this procedure kidney a. Does this mean that a should be ranked above any of the other kidneys in the set? Maybe not. Consider the case in which only a and some other kidney b are available, and assume that b was eliminated following the preliminary tests in the case of the choice problem, with ten kidneys. Given that there are only two kidneys in the new choice problem the doctors might now be able to run the detailed and expensive tests on both of them and discover that kidney b is actually better than a for patient x. Choice functions that can be derived from such procedures have been characterized by Manzini and Mariotti [20] and by Mandler [19].

[5]Choice functions that can be derived from such interactive procedures have been characterized by Xu and Zhou [26] and by Apesteguia and Ballester [3].

assumptions need to be made on the strategic behavior of agents in such mechanisms. Observe that the assumption of fully strategic rationality seems overly demanding in an environment in which individuals are not even individually rational. To resolve this tension, I define a notion of truthful implementation in Section 5. This notion can be viewed as a form of boundedly rational strategic behavior that approximates full strategic rationality.

Once all these hurdles are cleared, I adapt the First and Second Fundamental Theorem of Welfare Economics to the environment of house allocation problems with boundedly rational agents in Section 6. In accordance with the First Fundamental Theorem of Welfare Economics I find that any outcome of trade belongs to the larger set of Pareto optima. In accordance with the Second Fundamental Theorem of Welfare Economics, I find that any allocation in the smaller Pareto set is reachable through trade. Exchanging the two Pareto sets in the preceding two observations, one obtains stronger analogues of the Fundamental Theorems of Welfare. I show that these stronger analogues do not hold. These observations are all owed to the fact that the set of Pareto optima, according to the individual solid rankings of goods, can be very large whereas the smaller set according to the individual light rankings can be very small (it might even be empty). I show, in addition, that the sets of allocations that are implemented through trade differ for different trading mechanisms. This yields an interesting contrast to the the case of rationalizable behavior, in which these sets of allocations all coincide with the set of Pareto optima.

I first establish these results for *any* choice functions. Since these results all depend on the - potentially large - difference between the two Pareto sets, I go on to study cases in which this difference is smaller: I restrict the sets of permissible choice functions. In this context, I show that even when assuming that the agent's behavior only "minimally" deviates from rationalizable behavior, the main results of the article remain valid. Before going into detail, though, the notions of trade and of truthful implementation as well as the results are previewed with the help of a simple example (Section 3)

2 The Environment

The problems discussed in the article are represented by triplets $\mathcal{E} = (N, H, (c_i)_{i\in N})$, where N denotes the set of agents and H the set of objects that is to be matched to the agents; the vector $(c_i)_{i\in N}$ denotes the set of all agents' choice functions on H. The agents are simply numbered $N = \{1, ..., |N|\}$, it is assumed that there are equally many agents and objects: $0 \neq |N| = |H| < \infty$. In accordance with the convention adopted in the literature, the objects are called houses. Generic elements of the set of houses are denoted by x, y, w, and z. For each

$i \in N$, $c_i : \mathcal{P}(H) \to H$ with $c_i(S) \in S$ is a choice function representing the choice that agent i would make when given the opportunity to choose from a set S.[6] If a choice function c_i is rationalizable, I write $\succsim_i$ for the preferences that rationalize it. Note that the preferences $\succsim_i$ yield single-valued choice correspondences, if and only if $\succsim_i$ is a linear order, meaning that it is transitive, asymmetric and complete. Consequently, the present definition of a (house allocation) problem is standard, except that the agents' behavior is described by choice functions.[7]

For a particular problem $\mathcal{E} = (N, H, (c_i)_{i \in N})$, an **allocation** is defined as a bijection $\mu : N \to H$. Allocations are denoted by vectors μ, with μ_i denoting agent i's assignment. An allocation **rule** ϕ maps problems $\mathcal{E} = (N, H, (c_i)_{i \in N})$ into allocations $\phi(\mathcal{E})$.

To judge whether a particular allocation μ is Pareto-optimal or not, some notions of individual "preference" are needed. According to a the first and weaker definition, I say that agent i **lightly prefers** house x over house y, formally $xP_i^{\exists}y$, if there exists a set of houses $S \subset H$, such that $x, y \in S$ and $x = c_i(S)$. On the other hand, I say that agent i **solidly prefers** house x over house y, formally, $xP_i^{\forall}y$, if $y \neq c_i(S)$ for all $S \subset H$ with $x, y \in S$. Observe that $xP_i^{\forall}y$ holds if there exists no set, such that y is chosen when x is available; conversely, there needs to exist only one set containing x and y, such that x is chosen for $xP_i^{\exists}y$ to hold. The two notions of preference yield two notions of Pareto optimality: an allocation μ is called $P^{\forall}$***−Pareto-optimal*** ($P^{\exists}$***-Pareto-optimal***) if there exists no alternative allocation μ', such that $\mu_i' P_i^{\forall} \mu_i$ ($\mu_i' P_i^{\exists} \mu_i$ and $\mu_i' \neq \mu_i$) for all i in some non-empty subset K of the set of agents N and $\mu_i = \mu_i'$ for all other agents. I write $PO^{\forall}(\mathcal{E})$ ($PO^{\exists}(\mathcal{E})$) for the sets of $P^{\forall}$- ($P^{\exists}$-)Pareto-optimal allocations for problem $\mathcal{E}$. The notion of solid preference presented here ($P^{\forall}$) is identical with (or very similar to) the notions of preference that Bernheim and Rangel [8], Mandler [18], and Green and Hojman [15] use to compare outcomes in terms of individual and collective welfare.

Let me summarize some of the important properties of the two notions of preference. Note that $xP^{\forall}y$ implies $xP^{\exists}y$. Moreover, $yP^{\forall}x$ holds if and only if $xP^{\exists}y$ is violated. The relation $P^{\exists}$ is always complete, the relation $P^{\forall}$ need not be. The two relations coincide if and only if the underlying choice function is rationalizable which holds if and only if $P^{\exists}$ is transitive, which, in turn, holds if and only if $P^{\forall}$ is transitive. While both relations might violate transitivity, they do so in different ways. The statements $xP^{\forall}y$ and $yP^{\forall}z$ might hold, even if x and z are

[6] The set of all subsets of a set X is denoted by $\mathcal{P}(X)$.

[7] Note that the assumption of choice functions as a primitive of the model allows for a much larger range of "irrationalities" as the assumption of agents that maximize some intransitive and/or incomplete preferences. The latter assumption for example rules out an agent who chooses frozen yoghurt when offered ice-cream, frozen yoghurt, and broccoli, but does choose ice cream when only frozen yoghurt and ice-cream are available. Such choices are permissible in the present model. Examples of studies on trade and/or welfare that assume intransitive and/or incomplete preferences are Gale and Mas Collel [14] as well as Fon and Otani [13].

unranked by $P^\forall$. However, is is impossible that $zP^\forall x$ would hold for the given case. So, $P^\forall$ is acyclic. In contrast, for some choice functions $xP^\exists y$, $yP^\exists z$, and $zP^\exists x$ hold; the completeness of the relation $P^\exists$ implies that x and z must be ranked. In sum, these observations imply that the two different Pareto sets are nested for any problem: $PO^\exists(\mathcal{E}) \subset PO^\forall(\mathcal{E})$ holds for all problems $\mathcal{E}$. Moreover, for "highly irrational" choice functions there can be many cycles in $P^\exists$, and $P^\forall$ might leave many alternatives unranked. In this case, $PO^\exists(\mathcal{E})$ might be very small, even empty, and $PO^\forall(\mathcal{E})$ might be large.

With these definitions of, and observations on $P^\forall$- and $P^\exists$-Pareto optimality in hand, the guiding questions of the article can be formulated as follows: does free trade lead to $P^\forall$- ($P^\exists$-)Pareto-optimal allocations? Is there a way to allocate ownership rights, such that free trade results in a given $P^\forall$- ($P^\exists$-) Pareto-optimal allocation? Before a detailed definition and discussion of the concept of free trade in the given context, I will discuss these questions with the help of a simple example.

3 Example

Consider a house allocation problem $\mathcal{E}^* = (N, H, (c_i)_{i\in N})$ with three agents ($N = \{1, 2, 3\}$) and three houses $H := \{x, y, z\}$. Let the agents' choice functions be given by:

$$\begin{array}{llll} c_1(\{x,y,z\}) = x, & c_1(\{x,y\}) = x, & c_1(\{y,z\}) = y, & c_1(\{x,z\}) = z \\ c_2(\{x,y,z\}) = y, & c_2(\{x,y\}) = y, & c_2(\{y,z\}) = z, & c_2(\{x,z\}) = x \\ c_3(\{x,y,z\}) = z, & c_3(\{x,y\}) = y, & c_3(\{y,z\}) = z, & c_3(\{x,z\}) = x. \end{array}$$

The given house allocation problem has no $P^\exists$-Pareto optima. To see this, observe that, on the one hand, $(x, y, z) \notin PO^\exists(\mathcal{E})$ as (z, y, x) $P^\exists$-Pareto-dominates (x, y, z), since $c_1(\{x, z\}) = z$ and $c_3(\{x, z\}) = x$. On the other hand, any allocation $\mu \neq (x, y, z)$ is $P^\exists$-Pareto dominated by $(x, y, z) = (c_1(\{x, y, z\})), c_2(\{x, y, z\}), c_3(\{x, y, z\})$. The set of $P^\forall$-Paerto optima is non-empty. To calculate it, observe that $xP_1^\forall y$, $yP_1^\forall z$, $yP_2^\forall x$, $xP_2^\forall z$, $zP_3^\forall y$, and $yP_3^\forall x$. The set of $P^\forall$-Pareto optima contains three elements (x, y, z), (x, z, y) and (z, y, x).[8]

For now, let me identify the notion of free trade within the present environment with Gale's top trading cycles mechanism as defined by Shapley and Scarf [23]. In Section 4, this mechanism will be embedded in a larger class of trading mechanisms. According to the top trading cycles mechanism, each agent initially owns one house. The mechanism prescribes that each agent

[8]To determine the set of $P^\forall$-PO, all six possible allocations need to be checked. To see, for instance, that (y, x, z) is not $P^\forall$-Pareto-optimal, observe that agent 1 solidly prefers x to y, whereas agent 2 has the inverse $P^\forall$-preference.

points to "his most preferred" house and each house points to its owner. At least one cycle of agents and houses forms. All agents in these cycles are assigned the houses that they point to. The procedure is repeated with the remaining agents and houses until all agents have been assigned a house. The mechanism can be viewed as free trade from initial endowments, since, on the one hand, every house is owned by someone at any moment of the mechanism and, since, on the other hand, all exchanges are voluntary.

As far as the agents' behavior is concerned, I assume, for now, that at each stage each agent points to his choice out of the set of all remaining houses. If choices are rationalizable, this type of behavior coincides with truth-telling, which, in turn, is an equilibrium in the top trading cycles mechanism. This type of behavior could, therefore, be viewed as an approximation of strategically rational behavior. I discuss the agents' behavior in a mechanism and theories of implementation at length in Section 5.

Given this assumption on the agents' behavior, the top trading cycles mechanism implements the (unique) allocation $\mu = (x, y, z)$. To see this, observe that according to the hypothesis on the agents' behavior, each agent i points to $\mu_i = c_i(H)$ in the first stage of the mechanism - no matter which initial allocation they are starting out with. Absent any conflict of interest, each agent i is assigned μ_i in the first stage of the mechanism. In terms of the quest for welfare theorems for agents with choice functions that are not rationalizable, the following observations should be noted for this particular house allocation problem:

In tune with possible versions of the First and Second Fundamental Theorem, any outcome of free trade is $P^{\forall}$-Pareto-optimal and any $P^{\exists}$-Pareto-optimal allocation can be reached through free trade (in the case of the present example, the latter holds trivially, given that the set of $P^{\exists}$-Pareto optima is empty). These are the positive results. In terms of negative results, note that there are some $P^{\forall}$-Pareto optima that cannot be reached through free trade for *any* initial allocation. So when using the criterion of $P^{\forall}$-Pareto optimality, the Second Fundamental Theorem of Welfare Economics fails. Conversely, for the notion of $P^{\exists}$-Pareto optimality, the First Fundamental Theorem fails: there is an allocation that is reached for all initial allocations through free trade, even though this allocation is not $P^{\exists}$-Pareto-optimal. These four observations can conveniently be summarized as the subset relation $PO^{\exists}(\mathcal{E}^*) \subsetneq TR(\mathcal{E}^*) \subsetneq P^{\forall} - PO(\mathcal{E}^*)$, where the notation $TR(\mathcal{E}^*)$ stands in for the set of all allocations reachable through trade in the given housing problem $\mathcal{E}^*$. The main result of the article extends this subset relation to a much larger class of theories of trade and to all possible problems $\mathcal{E}$. In particular, I consider a class of trading mechanisms that allows for more unequal distributions of initial wealth than does the top trading cycles mechanism. I also consider implementation theories according to

which agents might not be so naive as to consider the set of *all* remaining houses as their actual choice set. Finally I show that the strictness of the subset relation remains valid when allowing only for "minimally irrational" problems $\mathcal{E}$.

4 Trading Mechanisms

In this section, I sketch out Papai's [21] definition of hierarchical exchange mechanisms, which constitutes a large superclass of Gale's top trading cycles mechanism. The two reasons for doing so correspond to the two motivations given in the introduction: first of all, one might criticize Gale's top trading cycles mechanism as too restrictive a notion of "free trade" as each agent starts out by owning exactly one house. In contrast, Papai's [21] hierarchical exchange mechanisms allow for the full spectrum of inequality of initial endowments, ranging from the most equal case, according to which each agent owns exactly one house, to the most unequal one, in which one single agent starts out owning all houses.

Secondly, to read the paper as an analysis of the behavioral welfare properties of mechanisms that have been suggested in the literature, it is useful to study a class of mechanisms that contains many of these mechanisms as special cases. This is the case for Papai's [21] hierarchical exchange mechanisms. Subsets of that class have been described by Abdulkadiroglu and Sonmez [2], Svensson [24], Ergin [12], Ehlers, Klaus, and Papai [11], Ehlers and Klaus [9], Kesten [17], Ehlers and Klaus [10], and Velez [25].

In a **hierarchical exchange mechanism**, all houses start out being owned by someone. In a first stage of the mechanism, the designer asks all agents to point to their most preferred houses. Each house points to its owner. At least one cycle of agents and houses forms. Any agent in such a cycle is assigned the house he points to and leaves the mechanism with this assignment (no agent or house can take part in two cycles). If an owner of multiple houses leaves the mechanism, the subset of his houses that have not been assigned is passed on to the agents who still await their assignments according to some fixed inheritance rule.[9] Agents are once again asked to point to their most preferred houses among the remaining ones and the same procedure is repeated, until each agents has been assigned a house.[10,11] A hierarchical

[9] So ownership rights in this class of mechanisms take two forms. Either an agent owns a house in the current period, or he faces the option to become an owner of a house.

[10] An explicit and detailed definition of hierarchical exchange mechanisms can be found in Papai [21] pp. 1408-1413.

[11] According to the definition by Papai [21], cycles are eliminated simultaneously. Bade [6] shows that for the standard case of rationalizable choice functions, the order of elimination does not matter. To see that, for the

exchange mechanism is denoted by Γ.

Hierarchical exchange mechanisms can be viewed as mechanisms arising out of the assignment of ownership and the ensuing free trades among owners. At any moment in the mechanism, each house is owned by someone, in the sense that the owner can appropriate the house as his assignment and leave the mechanism. Before the final assignments are made, the ownership of multiple houses is feasible. Any exchanges are voluntary. So one might argue that Papai's [21] hierarchical exchange mechanisms do represent the notion of trade that is appropriate for matching allocation problems. Since these mechanisms allow for a very broad and fine spectrum of initial endowments, ranging from maximal to minimal inequality, there are probably no other mechanisms for matching allocation problems that could be described as mechanisms of "free trade".[12] However, it is important to note that, as long as one identifies trade with some subclass of the set of hierarchical exchange mechanisms, the results of the present article apply.

One might argue, in addition, that hierarchical exchange mechanisms (or some subclass thereof) are of interest in their own right: most mechanisms that have been suggested in the literature as optimal - from some point of view or other - are comprised by the set of hierarchical exchange mechanisms. Some of these mechanisms have been put into practice in real-life matching allocation problems. It is, therefore, of interest to know how the sets of outcomes of such mechanisms relate to the sets of Pareto optima in the matching allocation problems in which they are used.

Hierarchical exchange mechanisms specify ownership rights and determine the agents who hold these rights. To answer the main questions of the article, it must be possible to assign these same rights in different ways to the different agents. If some mechanism Γ, for example, prescribes that agent 3 starts out owning houses h_1, h_4 and h_5, there should be an alternative mechanism that prescribes that agent 1 starts out owning these houses. To this end, I define a permutation $p : N \to N$ as an **initial assignment of ownership** for a hierarchical exchange

case of non-rationalizable choice functions, this order does matter, reconsider the Example provided in Section 3 together with Gale's top trading cycles mechanism with the initial endowment (x, y, z). In stage one, each agent points to his own endowment. According to the present definition of hierarchical exchange mechanisms, all three top trading cycles are eliminated simultaneously and (x, y, z) is the resulting assignment. If, on the other hand, only agent 2 and house y are eliminated in stage one, in stage two the agents 1 and 3 point to each other's houses. Given the sequential elimination of top trading cycles, the allocation (z, y, x) would be obtained.

[12] Pycia and Unver[22] characterize a superset of the set of all hierarchical exchange mechanisms: they show that the set of all strategy-proof and Pareto-optimal allocation mechanisms is the set of "trading cycles with brokers and owners". These mechanisms cannot be described as the result of ownership assignments and free trade alone. In Pycia and Unver [22]'s mechanisms, there are two types of control rights over houses: the rights of owners and the rights of brokers. A broker's control rights over some house *do not* involve the right to appropriate the house himself.

mechanism Γ. The mechanism that arises when permuting the roles of all agents by the bijection p is denoted by $p\Gamma$, and is also called a **permutation of a hierarchical exchange mechanism**. If under Γ some agent i is the initial owner of a subset of houses S, then for $p\Gamma$, agent $p(i)$ is the owner of this subset; if under Γ some agent i' is supposed to inherit house x after the initial owner i'' of x is assigned y, then under $p\Gamma$, agent $p(i')$ is to inherit x after $p(i'')$ is assigned y.

Hierarchical exchange mechanisms generalize the top trading cycles mechanism insofar as some agents might initially own multiple houses while others own none. Inheritance rules are introduced to solve the problem that agents may not own multiple houses in any *outcome* of a mechanism. Serial dictatorships constitute another special case of hierarchical exchange mechanisms: in this case, the first agent is the initial owner of all houses; he forms a cycle by pointing to one of his houses and is assigned that house; all remaining houses are inherited by the next dictator. Any permutation of serial dictatorship consists in a reordering of the sequence of dictators. Permutations of Gale's top trading cycles mechanism consist in permutations of the initial assignment of houses.

The definition of permuted hierarchical exchange mechanisms makes it possible to ask questions such as: is there an assignment of initial ownership rights, such that allocation μ arises for some given type of trading mechanism? The next section is concerned with the last missing link to answer such questions: of course, to know whether μ is implemented by some mechanism $p\Gamma$, we need to have a theory on the agents' behavior in a mechanism.

5 Implementation

To relate the allocations that can be reached via trade to the (different sets of) Pareto optima in some housing problem $\mathcal{E}$, we need to know which allocations are implemented by a mechanism. Standard notions of implementation presume individual rationality and can, therefore, not be applied directly to the present context. Moreover, the assumptions on the agents' strategic rationality, embodied in some notions of implementation, clash with the present assumption that agents are not even individually rational. In this section, I first define a notion of truthful implementation, which assumes that agents provide truthful answers to the designer's questions. In the ensuing discussion, I compare this notion of implementation with some other notions. I will argue, in particular, that the bounded strategic rationality associated with truthful implementation blends nicely with the assumption that agents are boundedly rational in terms of individual choice.

A mechanism Γ is said to **implement truthfully** an allocation μ in a housing problem $\mathcal{E}$ if μ is the allocation that results when all agents provide truthful answers to the designer. In

turn, an agent's response is considered ***truthful*** if the following two conditions are met. First of all, house x must be $P_i^{\forall}-$maximal in H' for agent i to claim that he likes house x most among the set of remaining houses H'. Secondly, there needs to be a set S' of at least two houses with $S \subset S' \subset H'$, such that $c_i(S') = x$, where the set S is the set of houses currently owned by agent i. Finally, T is called a ***theory of truthful implementation*** if it prescribes truthful answers to the designer's questions.

Let me first justify why such theories are called "truthful". Consider the first requirement: it implies that agent i cannot claim to like house y most out of the set H' when there exists some other house $z \in H'$, such that agent i would not choose y out of any subset S, if z was also available in that subset S. If preferences are rationalizable, this condition holds, if and only if agent i claims to most like the (unique) $\succsim_i$-maximal element in H'. However, given that agents' behavior is not necessarily rationalizable multiple houses might satisfy this requirement. This is where the second requirement comes into play. It states that agent i should have some theory on the set of houses S' he is actually choosing from, and that this theory should be consistent with the underlying facts. The houses currently owned by the agent (S) should certainly be in that imaginary choice-set S'. He should, moreover, not perceive any houses that already left the market to be part of his choice-set (which accounts for $S' \subset H'$). Finally, the requirement that S' should contain at least two elements implies that the agent should consider an actual choice situation when choosing to point to his own house. To see that the set of truthful theories of implementation is not empty, let $S' = H'$ for any set of remaining houses H'. Then $S \subset S' \subset H'$, $c_i(S') = c_i(H')$, and there is no other house in H' that is solidly preferred to $c_i(H')$.[13]

If the agents' behavior is rationalizable, hierarchical exchange mechanisms are strategyproof: in fact, truth-telling (announcing one's true preferences to the designer) is a dominant strategy (in the normal form representation of any hierarchical exchange mechanism). So we might ask: can truthful revelation as described above also be considered "equilibrium behavior" for the case of agents whose behavior is not rationalizable? A minimal requirement for a strategy profile to be an equilibrium should be that no agent can solidly improve upon his assignment when changing his strategy, holding everyone else's strategy fixed. It turns out that truth-telling need

[13]This is the theory of implementation used in the example in Section 3. One could also consider theories that have more of an equilibrium flavor. Such a theory could require that an allocation is implemented by a mechanism if the mechanism has an "e-strategy profile" that results in this allocation. A strategy profile is considered an "e-strategy profile" if at any stage the set S' that agent i based his choice $c_i(S')$ upon must not only contain all houses currently owned by the agent, but also all the houses which are "offered" to i according to the given strategy profile. Certainly these two theories are not the only ones that fit the notion of implementation advocated above.

not be an equilibrium as demonstrated by the following example:

Example 1 Let $H = \{x, y, z, w\}$. Let the choices of agents 1, 2, and 3 be rationalizable by $x \succ_i y \succ_i z \succ_i w$ for $i = 1, 2, 3$. Let agent 4's choice function be given by the conditions $c_4(S) \neq x$ if $z \in S$, $c_4(S) \neq w$ if $x \in S$, $c_4(S) \neq y$ if $w \in S$, $c_4(y, z, w) = w$, and $c_4(y, z) = y$. Together these choices imply that agent 4's behavior is not rationalizable and that $zP_4^{\forall}x$, $xP_4^{\forall}w$, and $wP_4^{\forall}y$ holds. Now let us consider the top trading cycles mechanism with an initial endowment $\mu = (x, z, w, y)$. Truth-telling implies that the agents with rationalizable preferences all point to house x. Agent 4 must point to house z since it is the unique maximizer of $P_4^{\forall}$ in H. There is only one cycle, and that cycle is of minimal length. Agent 1 leaves with his initial endowment x. In the next stage, agents 2 and 3 point to house y. Agent 4 must point to house w: he cannot point to y as $wP_4^{\forall}y$, he cannot point to z as $c_4(S') \neq z$ for any set S' containing z and at least one other element. In this stage, agents 3 and 4 exchange houses, and agent 2 keeps his endowment z. For this strategy profile, agent 2 would be better off by pointing to house y in the first stage of the mechanism. In that stage, agent 4 is willing to swap houses with him. If agent 2 chooses not to tell the truth in this stage, his final assignment is y instead of z, where $y \succ_2 z$.

In fact, the strategy profile described is the *unique* truthful strategy profile in the example. Consequently, the same example can be used to show that some hierarchical exchange mechanisms simply do not have truth-telling equilibria when the agents' behavior is not rationalizable. This observation forces a choice between truth-telling and equilibrium as solution concepts. Let me detail some reasons why I chose to focus on truth-telling.

First of all, the assumption that agents follow "equilibrium behavior" would place some stringent demands on the agents' ability to reason strategically. I view it as problematic to assume, on the one hand, that agents are not even individually rational, but to assume, on the other hand, that agents are strategically rational. If violations of rationality are due to the complexity of different decision situations, we should expect that strategic rationality is violated more often than individual rationality, given that strategic reasoning generally involves a high level of complexity.

Given the simple case of rationalizability, truth-telling is an equilibrium, therefore the assumption of truth-telling can be viewed as a first-order approximation to strategic behavior in the case of agents' behavior not being rationalizable. Agents who are not aware that their own behavior or some other players' behavior is not rationalizable might believe that by truthfully revealing their choices they are actually following equilibrium behavior.

Moreover, truthtelling could be interpreted as arising out of a form of shortsighted strategic

rationality: suppose agent i follows the truthful strategy and is assigned house μ_i. Suppose that H' is the set of houses still unassigned at the stage when agent i leaves the mechanism. The requirements of truthtelling imply that agent i will not regret having pointed to μ_i in that stage if he only considers that stage. There is no house left in H' that he would solidly prefer to μ_i; moreover there is an imaginary choice set S', out of which the agent would have chosen μ_i. In further illustration of this point, observe that the argument that agent 2 can obtain a solidly preferred house when deviating from the truthful strategy in Example 1 built on a comparison of the houses available to that agent in *different* stages of the mechanism.

Let me finally note that it is far from clear how the implementation through strategically rational behavior should be defined in the present context. Consider the definition according to which a strategy profile should be considered an equilibrium if no agent can obtain a house he solidly (or lightly) prefers to the one he is being assigned, if all agents follow the profile. Now observe that the top trading cycles mechanism implements *any* allocation according to this definition of equilibrium: simply fix the desired allocation as the initial endowment and assume the strategy profile according to which each agent points to their own initial assignment. According to this strategy profile, no player has any influence on the outcome of the mechanism and might as well opt to point to his own endowment in the first stage of the mechanism.

With the notion of truthful implementation in hand, we can now characterize the allocations that are implemented by hierarchical exchange mechanisms. I denote the outcome of a mechanism Γ with initial assignment of ownership p and the assumption of the theory of implementation T for a given house allocation problem $\mathcal{E}$ by $p\Gamma_T(\mathcal{E})$. Speaking in terms of the problem $\mathcal{E}$ defined in Section 3, we have that $p\Gamma_T(\mathcal{E}) = (x, y, z)$, where Γ is the top trading cycles mechanism, p is any permutation, and T is the theory of implementation according to which agents always point to their choice out of the set of all unassigned houses.

6 Welfare Theorems

Using the concepts developed in the prior two sections, the main observation of the introductory example can now be summarized as $PO^{\exists}(\mathcal{E}^*) \subsetneq \bigcup_p p\Gamma^*_{T^*}(\mathcal{E}^*) \subsetneq PO^{\forall}(\mathcal{E}^*)$, with $\mathcal{E}^*$ being the problem defined in that example, Γ^* the top trading cycles mechanism, and T^* the truthful theory of implementation according to which all agents at any stage point to their choice out of the set of houses remaining in the mechanism. In words: all $P^{\exists}-$Pareto optima of $\mathcal{E}^*$ are implementable through trade and any allocation that is truthfully implementable through trade is $P^{\forall}-$Pareto-optimal. These subset relations are strict: the allocation implemented through trade is *not* $P^{\exists}-$Pareto-optimal. Not every $P^{\forall}-$Pareto-optimal allocation is the outcome of

trade for some initial allocation of houses.

In this section, I show that these statements hold on a vastly more general level: namely, not just for identification of "trade" with the top trading cycles mechanism Γ^*, but with any hierarchical exchange mechanisms Γ, not just for T^*, but for all truthful theories of implementation, not just for the $\mathcal{E}^*$, but for all housing problems. The extensions of the First and Second Fundamental Theorem of Welfare Economics can now conveniently be stated as the following subset relations, where the intersection $\bigcap_{\Gamma,T}$ and union $\bigcup_{\Gamma,T}$ are to be understood over the entire set of hierarchical exchange mechanisms and the entire set of truthful theories of implementation.

Theorem 1 *Any $P^{\exists}$-Pareto optimum can be reached through any combination of a trading mechanism with a truthful theory of implementation. Any allocation that is truthfully implementable by some trading mechanism is $P^{\forall}$-Pareto-optimal. Thus,*

$$PO^{\exists}(\mathcal{E}) \subset \bigcap_{\Gamma,T}\left(\bigcup_{p} p\Gamma_T(\mathcal{E})\right) \text{ and } \bigcup_{\Gamma,T}\left(\bigcup_{p} p\Gamma_T(\mathcal{E})\right) \subset PO^{\forall}(\mathcal{E}) \text{ for all } \mathcal{E}.$$

The first subset relation translates to the following statement: for any $P^{\exists}$-Pareto-optimal allocation μ, any hierarchical exchange mechanism Γ, and any theory of implementation T, there exists an initial allocation of ownership p, such that μ is implemented by $p\Gamma$ for the theory of implementation T. This is a version of the Second Fundamental Theorem of Welfare Economics. The second subset relation extends the First Fundamental Theorem of Welfare Economics to the environment of house allocation problems without rationalizability: any outcome of trade is $P^{\forall}$-Pareto-optimal; this does not depend on any particular hierarchical exchange mechanism Γ, theory of implementation T, or initial assignment of ownership p.

Proof I start by showing that, for any $P^{\exists}$-Pareto-optimal allocation μ, there exists an ordering of agents, such that $\mu_i P_i^{\forall} \mu_k$ for all $k > i$ and all i. To see this, suppose for all agents i there existed some set $S_i \subset H$, such that $\mu_i \in S_i$ and $\mu_i \neq c_i(S_i)$. Now let all agents point to the owner of $c_i(S_i)$ under μ. Since there are only finitely many agents, at least one cycle forms. Consider the allocation μ' with $\mu'_i = c_i(S_i)$ for all the agents in the cycle and $\mu_i = \mu'_i$ for all other agents. Observe that all agents in the cycle lightly prefer their assignment under μ' to their assignment under μ: $\mu'_i = c_i(S_i) P_i^{\exists} \mu_i$ for all agents i in the cycle. This yields a contradiction with the $P^{\exists}$-Pareto optimality of μ. So there must be at least one agent who chooses μ_i whenever it is available. For this agent we have that $\mu_i P_i^{\forall} x$ for all $x \in H$. Let $i = 1$. Since the restriction of the allocation μ to the subsets of all remaining agents $\{2, \cdots, |N|\}$ and houses $H \setminus \{\mu_1\}$ has to be also $P^{\exists}$-Pareto-optimal, the conjecture follows by induction.

Next, I show that for any $P^{\exists}$-Pareto-optimal μ and any mechanism Γ and theory of implementation T, there exists an initial assignment of roles p, such that μ is the outcome of $p\Gamma_T$. By the first paragraph of the proof, the $P^{\exists}$-Pareto optimality of μ allows me to order the agents, such that $\mu_i P_i^{\forall} \mu_k$ for all $k > i$ and all i. Now define p, such that at any given stage an agent i owns a house if any agent $l < i$ either currently owns a house or already left the mechanism with his assignment. Assume, furthermore, that any agent i who is an owner of houses (also) owns house μ_i. Now observe that for any theory of implementation T, agent i must point to houses with indices $l \leq i$, since $\mu_i P_i^{\forall} \mu_k$ for $k > i$. This means that at any stage (at least) the agent with the lowest index leaves the mechanism. Furthermore, no agent's request for a house with an index lower than his own will ever be granted; this would require for at least some other agent to accept a house with an index higher than his own. Consequently, the allocation μ arises out of the mechanism Γ for the given theory of implementation T.

To see that every outcome of some $p\Gamma_T$ is $P^{\forall}$-Pareto-optimal, let μ be such an outcome and assume w.l.o.g. that agents $\{1, \cdots, l\}$ have been assigned $\{\mu_1, \cdots, \mu_l\}$ in the first stage of the mechanism. So each of these agents i must have pointed to μ_i in the first stage, which implies that μ_i is $P_i^{\forall}$-optimal in H for each i (by the requirement of the theory of implementation T). By the same argument, the houses assigned in the second stage must be $P_i^{\forall}$-optimal in $H \setminus \{\mu_1, \cdots, \mu_l\}$ for the agents. Proceeding inductively, we can conclude that μ is $P^{\forall}$-Pareto-optimal. □

The next theorem summarizes the negative results of the article:

Theorem 2 *An allocation might not be $P^{\exists}$-Pareto-optimal, even if it is implementable by any combination of a trading mechanism with some truthful theory of implementation. Some $P^{\forall}$-Pareto-optimal allocations cannot be truthfully implemented through any trading mechanism. Different trading mechanisms truthfully implement different sets of allocations. Formally, these statements can be expressed as follows: there exist house allocation problems $\mathcal{E}', \mathcal{E}''$, and $\mathcal{E}'''$ and hierarchical exchange mechanisms $\widetilde{\Gamma}$ and $\overline{\Gamma}$, such that*

$$PO^{\exists}(\mathcal{E}') \not\supseteq \bigcap_{\Gamma,T} \left(\bigcup_{p} p\Gamma_T(\mathcal{E}') \right); \qquad \bigcup_{\Gamma,T} \left(\bigcup_{p} p\Gamma_T(\mathcal{E}'') \right) \not\supseteq PO^{\forall}(\mathcal{E}'')$$
$$\text{and } \bigcup_{p,T} p\widetilde{\Gamma}_T(\mathcal{E}''') \neq \bigcup_{p,T} p\overline{\Gamma}_T(\mathcal{E}''').$$

The first statement implies a failure of the First Fundamental Theorem of Welfare Economics when adopting the notion of $P^{\exists}$-Pareto optimality. It posits the existence of some housing problems and allocations μ with the following features: for any notion of trade, there exists an allocation of ownership rights, such that μ arises as the outcome of trade for the given house

allocation problem. Still μ is not $P^{\exists}$-Pareto-optimal in that problem. The second statement shows the limits of the Second Welfare Theorem when adopting $P^{\forall}$-Pareto optimality as the measure of welfare: there are house allocation problems with $P^{\forall}$-Pareto-optimal allocations that cannot be reached by *any* hierarchical exchange mechanism Γ, initial allocation of ownership p, and theory of truthful implementation T. Finally, the last inequality implies that different mechanisms of hierarchical exchange generally yield different sets of allocations for non-rationalizable choice functions.

In the case of rationalizable preferences the full analog of both Fundamental Theorems of Welfare Economics holds: in that case, we have $P_i^{\exists} = P_i^{\forall}$ and, therefore, $PO^{\exists}(\mathcal{E}) = PO^{\forall}(\mathcal{E}) := PO(\mathcal{E})$. Abdulkadiroglu and Sonmez [1] showed that serial dictatorship as well as Gale's top trading cycles mechanism each trace out the full Pareto set of any house allocation problem, in terms of the formalism presented here $\bigcup_p p\widetilde{\Gamma}(\mathcal{E}) = \bigcup_p p\overline{\Gamma}(\mathcal{E}) = PO(\mathcal{E})$[14] for $\widetilde{\Gamma}$ and $\overline{\Gamma}$ Gale's top trading cycles mechanism and serial dictatorship respectively. Papai [21] showed that hierarchical exchange mechanisms are Pareto-optimal, in the sense that $\bigcup_\Gamma \left(\bigcup_p p\Gamma(\mathcal{E})\right) \subset PO(\mathcal{E})$. Bade [6] complements this version of the First Fundamental Theorem of Welfare Economics with a version of the Second to conclude that $\bigcup_p p\Gamma(\mathcal{E}) = PO(\mathcal{E})$ holds for any hierarchical exchange mechanism Γ. The next three examples of house allocation problems make up the proof of Theorem 2.

Proof To see that the outcomes of different hierarchical exchange mechanisms need not coincide, consider the following example:

Example 2 Let $H = \{x, y, z\}$, and let the choices of agents 1 and 2 satisfy $c_1(\{x, y, z\}) = x$, $c_1(\{x, z\}) = z$, $c_2(\{x, y, z\}) = y$, and $c_2(\{y, z\}) = z$. Let agent 3's behavior be rationalizable by $x \succ_3 y \succ_3 z$. Observe that $\mu = (x, y, z)$ is achievable under top trading cycles. To see this, assume μ as the initial allocation and assume a theory of implementation according to which agents point to $c_i(H)$ in the first stage of the mechanism. The allocation (x, y, z) already obtains in the first (and therefore last) stage of the mechanism. To see that μ is not implementable via serial dictatorship, observe that under serial dictatorship either agent 1 or agent 2 has to be the first dictator for (x, y, z) to result. If agent 1 is the first dictator, neither one of the two remaining agents would pick μ_i as the second dictator. The same holds for the case in which agent 2 is the first dictator.

[14] Note that I am not mentioning any theory of implementation. In the case of all choice functions being rationalizable, there is a unique theory of truthful implementation. This theory corresponds to the play of dominant strategies in the normal form game representation of the mechanism.

To see that not every $P^{\forall}$-Pareto-optimal allocation is reachable through some hierarchical exchange mechanism, consider the following example:

Example 3 Let $H = \{x, y, w, z\}$ and let the choice functions of agents 1, 2 and 4 be rationalizable by $x \succ_i y \succ_i z \succ_i w$ for $i = 1, 2, 4$. Let the choice function of agent 3 be such that $c_i(S) = y$, if $y \in S$, $c_3(\{x, z, w\}) = z$, $c_3(\{z, w\}) = w$. The allocation $\mu = (x, y, z, w)$ is $P^{\forall}$-Pareto optimal. However, there is no hierarchical exchange mechanism and implementation theory that implements μ. To see this, observe that when all houses are still available, x is the unique $P_i^{\forall}-$maximizer for agents i=1,2, and 4, and y is the unique $P_3^{\forall}-$maximizer. Therefore, in the first stage of any hierarchical exchange mechanism implementing μ, (only) agent 1 and house x will be matched. By the same logic, in the second stage only agent 2 and house y will be matched. In the third stage, the two remaining agents must view $\{z, w\}$ as their choice set. Since $c_3(\{z, w\}) = w$ and $c_4(\{z, w\}) = z$, the resulting allocation cannot be μ.

To see that some allocations are not $P^{\exists}$-Pareto-optimal, even though they can be reached through any hierarchical exchange mechanisms, consider the following example.

Example 4 Let $H = \{x, y, z, w\}$. Let the choices of agents 1 and 2 be rationalizable by $x \succ_i y \succ_i z \succ_i w$ for $i = 1, 2$. The choice functions of agents 3 and 4 have the following features: $y \in S$ implies that $c_3(S) = y$, $c_3(\{x, z, w\}) = w$, $c_3(\{z, w\}) = z$, $x \in S$ implies $c_4(S) = x$, $c_4(y, z, w) = z$ and $c_4(\{z, w\}) = w$. Observe that $\mu = (x, y, z, w)$ is not $P^{\exists}$-Pareto-optimal as $wP_3^{\exists}z$ and $zP_4^{\exists}w$, since $c_3(\{x, z, w\}) = w$ and $c_4(y, z, w) = z$. However, μ is reachable for any theory of implementation and any hierarchical exchange mechanism. To see this, fix a hierarchical exchange mechanism and a theory of implementation. Define p, such that agent 1 is the initial owner of house x and house y is not initially owned by agent 3. In the first stage of the mechanism, when no house has been assigned yet, there exist unique maximizers of all agents $P_i^{\forall}-$rankings. Agents 1, 2, and 4 point to house x; agent 3 points to house y. For the given assumption on the initial assignment, only house x and agent 1 leave the market. Define p, such that agent 2 inherits house y, if he does not already own it. If there are two owners in the second stage, define p, such that agent 3 is the other owner of a house. If there are 3 owners, define p, such that agent 4 owns house w in the second stage. The requirement that agents should point to houses that are $P_i^{\forall}-$maximal among the remaining houses implies that agents 2 and 3 must both point to house y. Given that agent 2 owns house y, he leaves the market with this house. If agent 4 owns w and points to it for the given theory of implementation, he appropriates this house and the desired allocation obtains. If not, there is a third stage, with only agents 3 and 4 and houses z and w remaining. In this stage both must consider $\{z, w\}$ as

their choice set. The last two agents' choices from this set are such that the desired allocation obtains.

□

The intuition for the difference between the sets of allocations implemented by different mechanisms is that different mechanisms focus the agents on different choice sets when making their decisions. If the agents' behavior is rationalizable, such focus - and, therefore, the choice of a particular hierarchical exchange mechanism - does not matter. However, in the present case, such focus does lead to different choices and thereby different sets of outcomes. A similar intuition can explain the gaps between the two Pareto sets and the set of allocations that are implementable through trade. The reason why the allocation μ in Example 3 is not implementable is that at the stage when agent 3 needs to choose house z to obtain the allocation μ, such a choice is not compatible with truth-telling. The house z is $P_3^{\forall}$-optimal among the remaining houses; however, no choice set supports agent 3's choice of z out of the remainder. The $P_3^{\forall}$-optimality of z is owed to the comparison of z with houses that are long gone and should therefore not have an effect on choices in the current stage. The gap between the set of $P^{\exists}$-Pareto optima and the set of allocations that are implementable through any mechanism can be explained using the same logic. In fact, the allocation μ in Example 4 is such that agents 3 and 4 can $P^{\exists}$-improve by exchanging their assignments. However, once houses x and y have left the market, both agent 3 and agent 4 would always choose in accordance with μ. Since for all mechanisms there exist allocations of initial wealth p, such that x and y leave the market before the other two houses, the allocation μ can be obtained for any hierarchical exchange mechanism.

Note that Theorems 1 and 2 remain valid, if we restrict our attention to any subsets of the set of all hierarchical exchange mechanisms or the set of all theories of truthful implementation, as the intersection of a smaller set of sets, is a superset of the intersection of a larger set of sets and as the union of a smaller set of sets is a subset of the union of a larger set of sets. This observation implies that the results do not depend on the permissive interpretation of "trade" as the set of *all* hierarchical exchange mechanisms. If instead one wants to consider only top trading cycles or any other subset of hierarchical exchange mechanisms as the appropriate set of trade mechanisms, the results remain valid. The same holds for the set of theories of truthful implementation.

There are already some versions of the First and Second Fundamental Theorem of Welfare Economics in the literature on agents with non-rationalizable choices. All results that I am aware of concern market environments with divisible goods. Bernheim and Rangel [8] prove

a First Fundamental Theorem of Welfare Economics for markets that are standard except for the assumption that the agents' behavior need not be rationalizable. Their notion of Pareto optimality relies on a notion of preferences that is very similar to the $P^{\forall}$-preferences defined in the present article. This result lies very much in line with Theorem 1. Interestingly, Mandler [18] proves a version of the Second Theorem of Welfare Economics that also defines Pareto optimality with respect to $P^{\forall}$-preferences. This result stands in sharp contrast with Theorem 2 of the present article, which claims that the Second Fundamental Theorem of Welfare Economics does not apply to the case non-rationalizable agents when using the notion of $P^{\forall}$-Pareto optimality. Of course, there are some major differences in the framing of the problem: Mandler [18] studies quasi-equilibria in a market environment that is standard except for the assumption that the agents' choices need not be rationalizable, whereas I study the outcomes of hierarchical exchange mechanisms that match some indivisible goods to agents. But these differences do not drive the stark discrepancy of the results. This discrepancy is caused by my imposition that for any agent's choice in a mechanism there needs to be some set that is consistent with the underlying facts, such that the agent's choice can be construed as a choice from this set. The imposition of such a condition on Mandler's [18] trading environment would possibly also considerably shrink the set of allocations that are implementable through trade in his environment. The same arguments apply to the difference between my results and the ones by Fon and Otani [13], who prove a version of the First Fundamental Theorem and some price characterizations of the set of Pareto optima, based on the assumption of intransitive and/or incomplete preferences. The assumption of such preferences rules out many 'irregularities" that are permissible under the present framework. In their approach, just like in Mandler's, individuals are always willing to select any preference-maximal element of a choice set. An additional condition has to be satisfied in my framework: the preference-maximal element has to be chosen out of a set S that is consistent with the agent's understanding of the mechanism.

7 Minimally Irrational Behavior

Up until now, *any* deviations from rationalizability were permitted. In the present Section I ask how the main two theorems of the paper would change, if we limited our interest to particularly appealing or small deviations of the assumption of rationalizability. Theorem 1 shows that the set of all outcomes of trade is nested between the two Pareto sets for *any* problem $\mathcal{E}$. It consequently holds unchanged for any subset of problems.

In contrast, Theorem 2 makes three existence claims. Some problems have allocations that can be reached through any trading mechanism and truthful theory of implementation, even

if they are not $P^{\exists}$-Pareto-optimal. Other problems have $P^{\forall}$-Pareto optima that cannot be reached through any combination of trading mechanism with a truthful theory of implementation. Finally, for some problems, the sets of allocations that are truthfully implementable through different mechanisms might differ. So the question is whether such problems exist when we limit the set of permissible problems.

Theorem 2 fails when restricting attention to problems with rationalizable choice functions. In Section 2, I noted that the subset relation $PO^{\exists}(\mathcal{E}) \subset PO^{\forall}(\mathcal{E})$ holds for *all* $\mathcal{E}$. If all agents' choice functions are rationalizable, we have that $PO^{\exists}(\mathcal{E}) = PO^{\forall}(\mathcal{E})$. It is to be expected that the gap between the two Pareto sets $PO^{\forall}(\mathcal{E}) \setminus PO^{\exists}(\mathcal{E})$ increases with the "degree" of irrationality present in the problem $\mathcal{E}$. One might conjecture that Theorem 2 fails (or at least some parts of it do) if we restrict attention to problems with minimally irrational behavior. The question is: how irrational do the agents need to be for Theorem 2 to hold? The answer is: not much at all!

To see this, note that all three examples used in the preceding proof either posit that an agent's choices are rationalizable or that they can be (up to renaming of alternatives) represented by a choice function c_i on $H = \{x, y, z\}$ with $c_i(H) = x$ and $c_i(\{x, y\}) = y$. None of the examples makes any further assumptions on the choices of agents from the sets $\{x, z\}$ and $\{y, z\}$. Now observe that, to be non-rationalizable at all, some violation of the weak axiom of revealed preference must exist. Since the choice function c_i posits *nothing* beyond a single such violation, one must consider the difference between c_i and a rationalizable choice function minimal. In fact, since c_i does not specify choices from the sets $\{x, z\}$ and $\{y, z\}$ these can be determined, such that they represent a minimal deviation from rationalizability according to any theory that measures the degree of such a deviation.

This means, in particular, that Theorem 2 continues to hold when restricting attention to behavior that is sequentially rationalizable by just two rationales as defined by Manzini and Mariotti [20], or to behavior that can be explained as choices via checklist of length two as defined by Mandler [19]. Nothing changes when considering only behavior that is rationalizable by a game tree with just two agents and two nodes as defined by Xu and Zhou [26]. By the same logic, Theorem 2 remains valid when considering decision makers whose behavior can be rationalized with at most two rationales following Kalai, Rubinstein and Spiegler [16], or when considering decision makers that are minimally irrational following Ambrus and Rozen [5] or following Apesteguia and Ballester [4]. No matter how little irrationality we permit in housing problems, there are always some $P^{\forall}-$Pareto-optimal allocations that are not truthfully implementable by any hierarchical exchange mechanism. Different mechanisms will generally implement different sets of allocations, even if we only allow for behavior that minimally deviates

from rational behavior.

8 Conclusion

This paper set out to shed some light on the relation between Pareto optimality and trade under the assumption that agents are boundedly rational. The problem was framed in an environment of matching allocation problems. I defined two nested sets of (behavioral) Pareto optima (the sets of $P^{\forall}-$ and $P^{\exists}-$Pareto optima) for any matching allocation problem. Given that one uses the more encompassing notion of Pareto optimality, the statement that any outcome of trade is Pareto-optimal holds true, whereas the statement that any Pareto-optimal outcome can be reached through trade does not. If one uses the more restrictive notion of Pareto optimality, one obtains that any Pareto optimum can be arrived at through trade. However, the complementary statement that any outcome of trade is Pareto-optimal does not hold for the more restrictive notion of Pareto optimality.

While these results were first obtained for *any* non-rationalizable choice functions, they were considerably strengthened by showing that the results do not change when admitting only for "minimal deviations" from rationalizability. This latter observation implies, in particular, that the results are also valid for the environments of kidney exchanges and school allocation, as described in the introduction. Even if all agents' choices follow decision procedures that sequentially eliminate options like the doctors' choices described in the introduction, there are some $P^{\forall}$-Pareto optima that cannot be reached through trade. Even if we assume that schooling choices are determined by the strategic interplay of only two agents, who each act as maximizers of transitive and complete preferences, different mechanisms of trade lead to different sets of outcomes.

This last observation opens up some new questions on the design of matching mechanisms. For rationalizable choice functions, the sets of allocations implemented do not differ across different trading mechanisms. Any question that relies on the comparison of the outcome sets of different mechanisms (for all initial allocations of ownership) is therefore idle. This changes dramatically once we relax the assumption of rationalizability. We could now pose such questions as: which mechanism leads to more egalitarian allocations? Or, which mechanism is more indeterminate, in the sense that it has a larger set of outcomes? Or, which mechanism privileges choices from large sets, in the sense that more agents receive houses they would choose out of larger sets?

Another interesting arena for research is to take the intuitions behind the introductory stories about decision procedures seriously. One could, for example, assume that patients do

have (linear) preferences over kidneys, but that it is costly to learn theses preferences. An allocation mechanism would then interact with some form of strategic information acquisition. One could go on to use the agents' ex ante utilities for different mechanisms to Pareto-rank different mechanisms. In Bade [7], I provide some preliminary observations on this topic and show, in particular, that not every Pareto Optimum can be reached through free trade in the case that agents can endogenously acquire information about the objects to be distributed. Similarly, one could explicitly model the interaction between family members when selecting a mechanism for school choice. In any case, the results on the connection between trade and Pareto optimality derived in this article will possibly still be of use as starting points for such studies.

References

[1] Abdulkadiroglu, A. and T. Sonmez: "Random Serial Dictatorship and the Core from Random Endowments in House Allocation Problems",*Econometrica*, 66, (1998), pp. 689-701.

[2] Abdulkadiroglu, A. and T. Sonmez: "House Allocation with Existing Tenants", *Journal of Economic Theory*, 88, (1999), pp. 233-260.

[3] Apesteguia J. and M. Ballester: "Choice by Sequential Procedures", *mimeo*, Universidad Autónoma de Barcelona, 2010.

[4] Apesteguia J. and M. Ballester: "A Measure of Rationality and Welfare", *mimeo*, Universidad Autónoma de Barcelona, 2010.

[5] Ambrus A. and K. Rozen: "Rationalizing Choice with Multi-Self Models", *mimeo*, Harvard, 2010.

[6] Bade, S. "Pareto-Optimal Assignment by Hierarchical Exchange", *mimeo*, MPI for Research on Collective Goods, Bonn, 2010.

[7] Bade, S. "Matching Allocation Problems with Endogenous Information Acquisition", *mimeo*, MPI for Research on Collective Goods, Bonn, 2010.

[8] Bernheim D. and A. Rangel: "Beyond Revealed Preference: Choice Theoretic Foundations for Behavioral Welfare Economics", *Quarterly Journal Of Economics*, forthcoming.

[9] Ehlers L. and B. Klaus: "Resource-Monotonicity for House Allocation Problems", *International Journal of Game Theory*, 32, (2004), pp. 545-560.

[10] Ehlers L. and B. Klaus: "Consistent House Allocation", *Economic Theory*, 30, (2007), pp. 561-574.

[11] Ehlers L., B. Klaus and S. Papai: "Strategy-Proofness and Population-Monotonicity for House Allocation Problems", *Journal of Mathematical Economics*, 38, (2002), pp. 329-339.

[12] Ergin, H.: "Consistency in House Allocation Problems", *Journal of Mathematical Economics*, 34, (2000), pp. 77-97.

[13] Fon, V. and Y. Otani: "Classical Welfare Theorems with Non-transitive and Non-complete Preferences", *Journal of Economic Theory*, 20, (1979), 409-418.

[14] "An Equilibrium Existence Theorem for a General Model without Ordered Preferences," *Journal of Mathematical Economics*, 2, (1975), 9-15.

[15] Green, J. and D. Hojman: "Choice, Rationality and Welfare Measurement", *mimeo* Harvard University, (2008).

[16] Kalai, G. A. Rubinstein and R. Spiegler: "Rationalizing Choice Functions by Multiple Rationales", *Econometrica*, 70, (2002), pp. 2481-2488.

[17] Kesten O.: "Coalitional Strategy-Proofness and Resource Monotonicity for House Allocation Problems" , *International Journal of Game Theory*, 38, (2009), pp. 17-22.

[18] Mandler, Michael.: "Indecisiveness in behavioral welfare economics Michael Mandler", *mimeo*, Royal Holloway College, University of London (2010).

[19] Mandler, Michael.: "Rational Agents are the Quickest", *mimeo*, Royal Holloway College, University of London (2010).

[20] Manzini, P. and M. Mariotti: "Sequentially Rationalizable Choice", *American Economic Review*, 97, (2007), pp. 1824-1839.

[21] Papai, S.: "Strategyproof Assignment by Hierarchical Exchange", *Econometrica*, 68, (2000), pp. 1403-1433.

[22] Pycia, M. and Unver, U.: "A Theory of House Allocation and Exchange Mechanisms," *mimeo* UCLA, January 2009.

[23] Shapley, L. and H. Scarf: "On Cores and Indivisibility", *Journal of Mathematical Economics*, 1, (1974), pp. 23-37.

[24] Svensson L. G.: "Strategy-proof allocation of indivisible goods", *Social Choice and Welfare* 16, (1999), pp. 557-567.

[25] Velez, R.: "Revisiting Consistency in House Allocation Problems and the Computational Approach to the Axiomatic Method", *mimeo*, (2007), University of Rochester.

[26] Xu, Y. and L. Zhou: "Rationalizability of Choice Functions by Game Trees", *Journal of Economic Theory*, 134, (2007), pp. 548-556.