Irrationality-proofness: markets versus games*

(forthcoming in the *International Economic Review*)

Michael Mandler†

Royal Holloway College, University of London

This version: December 2013

Abstract

How robust are economic models to the introduction of irrational agents? The Pareto efficiency of competitive equilibria is not robust: one irrational agent leads to inefficiency. But the property that rational agents cannot use their own resources to Pareto-improve on their competitive allocation holds regardless of the number of irrational agents. Full production efficiency can be robust as well, but irrational firms introduce a trade-off between efficiency and the attainment of Pareto improvements. Regarding games, we show that while existing implementation mechanisms are sensitive to the presence of irrational agents there are robust alternatives with attractive welfare properties.

JEL codes: D01, D51, D61, D78, C72.

Keywords: irrationality, Pareto efficiency, general equilibrium, implementation, production efficiency.

*I thank Hanming Fang, Sophie Bade, and two referees for several helpful comments and suggestions.

†Address: Department of Economics, Royal Holloway College, University of London, Egham, Surrey, TW20 0EX, United Kingdom. Email: m.mandler@rhul.ac.uk

1 Introduction

We consider properties of economic equilibria that are 'irrationality-proof,' that is, robust to the inclusion of irrational agents. Irrational agents are simply consumers who make errors when maximizing utility subject to a budget constraint and firms that do not maximize profits. A property's degree of irrationality-proofness is gauged by the number of irrational agents who can be added to a model without overturning the property.

If just one consumer or firm in a general equilibrium model chooses irrationally, a competitive equilibrium need not be Pareto efficient. The Pareto efficiency of competitive equilibria thus exhibits the lowest level of irrationality-proofness and its account of the welfare advantages of a market economy must therefore be misleading. As a replacement for Pareto efficiency, we show that, regardless of the number of irrational agents, the rational consumers and firms cannot Pareto improve on their equilibrium allocation if they are restricted to use only their own endowments and technologies, a property we call 'Pareto efficiency for the rational agents.' Since this property prevails no matter how many agents are irrational, it exhibits the highest level of irrationality-proofness. The proof that competitive equilibria enjoy this property is only a tiny variant of the classical argument that competitive allocations are in the core, but the applicability of the classical argument to models with irrational agents does not seem to have been noticed.

More important results hold for production economies. If the rational firms in the aggregate have a production set that contains the production set of the irrational firms, and if production sets satisfy a limited constant-returns property then full production efficiency obtains. The production efficiency of competitive equilibria thus displays an intermediate degree of irrationality proofness: it can persist in the presence of some irrational firms but not if there are so many irrational firms that the rational firms' technology fails to dominate that of the irrational firms. The contrast between full efficiency on the production side and 'Pareto efficiency for the rational' on the consumer side supports the folk wisdom that competitive markets root out inefficiency in production while leaving irrational households untouched. Error-prone consumers have the room to persist in their mistakes, but markets do not grant firms the same leeway.[1] Becker (1957) argued long

[1]Evolutionary selection, as in Sandroni (2000) and Blume and Easley (2006), would qualify this

ago that an attempt by irrational firms to racially discriminate in hiring can be made irrelevant by competition from fully rational firms.

Full production efficiency can obtain in the presence of irrational firms due to creative destruction: the rational firms drive the irrational out of business. This causal mechanism for production efficiency leads to distinctive policy conclusions. Economists have customarily turned to lump-sum payments to compensate agents that would be harmed by and therefore might block policy reforms; lump-sum payments to rational agents do not lead to inefficiency. But an irrational firm that receives a compensation payment can make inefficient decisions and remain shielded from bankruptcy. Irrational firms can therefore jeopardize the long tradition, based on the second welfare theorem, of using compensation payments to design Pareto improvements. Consider the conversion of crop subsidies into lump-sum payouts in the Common Agricultural Policy of the EU or the granting of carbon permits to firms to mitigate the burden of a carbon tax. Normally economists back these policies, but the presence of irrational firms can overturn this advice.

To see if markets are unusually robust to the addition of irrational agents, we compare markets to games that fully implement competitive allocations. If everyone is rational, implementation games can closely approximate the outcomes of competitive markets; in some cases, their equilibrium outcomes can exactly coincide with the competitive equilibrium outcomes. But if some agents are irrational then the most well-known full implementation games can have no equilibria or have equilibria that are not Pareto efficient for the rational agents. For example, the famous Hurwicz-Maskin-Postlewaite (1995) implementation game typically has no equilibria when just a single agent is irrational: existence of equilibrium in this game therefore fails to show even the lowest degree of irrationality-proofness. This fragility may be one reason why the equilibria of Nash implementation games can seem implausible. But there are alternative games where irrational agents do less damage: we construct games whose outcomes are Pareto efficient for the rational agents at every equilibrium, regardless of how many irrational agents are present. The 'Pareto efficiency for the rational' conclusion for games echoes our results for competitive markets, even though the formal arguments at work in the two settings have little

conclusion. As we explain in section 3, our arguments have no evolutionary component.

in common. The parallelism suggests that Pareto efficiency for the rational will hold widely in models with irrational agents. On the other hand, in the game setting Pareto efficiency for the rational does not imply that the rational agents will usually gain from the presence of irrational agents, as they do in competitive equilibria.

We assume that rational agents in games best-respond to the actions taken by the irrational agents. This approach stakes out a middle ground that avoids both the extreme rationality assumption that all agents play best responses and the position that the consequences of irrationality are so unforeseeable that rational agents must adopt actions that are always optimal regardless of how irrational agents play. The latter approach would amount to a dominant strategy requirement that we show leads an to impossibility result.

In both the market and game settings, the agents in this paper make errors: consumers who fail to solve constrained maximization problems, firms that do not maximize profits, and players who fail to best-respond. Prominent among the sources of error are the rules of thumb that arise when agents, out of inertia, stick to old decision rules that have lost their validity. But irrationality does not entail unpredictability. If anything, agents who follow rules of thumb are easy to predict. Equilibrium analysis is therefore suitable: rational agents should be able to adjust their actions to the behavior of the irrational agents in such a way that a profile of mutually consistent actions can emerge.

In our results, Pareto efficiency for the rational agents illustrates the highest level of irrationality-proofness; it holds no matter how many irrational agents are present. Other properties of market equilibria exhibit the next best level: they hold when just a single agent is rational. No-arbitrage conditions in finance typically display this degree of irrationality-proofness. If an arbitrage opportunity is present – for example, if asset prices fail to satisfy a martingale – then every agent must be failing to exploit an opportunity to make a risk-free profit and hence must be irrational.[2]

For some phenomena, it has long been the norm to consider the effect of irrational agents, for example, the impact of noise traders on financial markets (see, e.g., De Long

[2]The martingale property of asset prices and its link to rationality requirements originates in Samuelson (1965); the modern approach begins with Harrison and Kreps (1979). The strong irrationality-proofness of the property is well-known, but it is difficult to document an explicit statement.

et al. (1990)). The impact of partisan voters – agents who always vote for the same candidate regardless of their information – on information aggregation (Feddersen and Pesendorfer (1996)) can also be understood as an analysis of irrationality-proofness. But general theories of the effect of irrational agents have been rare. Haltiwanger and Waldman (1985, 1989) consider various games with the express purpose of seeing how irrational agents affect equilibrium outcomes; their results turn on whether strategic substitutes or complements are present. Fehr and Tyran (2005) and Camerer and Fehr (2006) have deepened this line of analysis. Sutton (1997) analyzes a class of industrial organization games with a 'one rational agent is enough' degree of irrationality-proofness, comparable to the no-arbitrage conditions of finance.

In contrast to the above literatures, when we consider the irrationality-proofness of games we take the implementation point of view: we do not consider the effect of irrational agents on a specific game but on what games with irrational agents can in principle achieve. Our analysis is therefore related to Eliaz (2002), which we discuss in section 4. We do however share one feature with the above literatures: our agents are either rational or irrational. Another way to introduce a small amount of irrationality is to let agents – possibly all agents – be a little irrational; the quantal response equilibria of McKelvey and Palfrey (1995) is a leading case in point.

2 Irrational consumers

The analysis of exchange economies with irrational agents is straightforward, involving only a simple variation on Lloyd Shapley's proof that competitive allocations lie in the core. But this argument is the natural place to start and it shows how well-suited the theory of the core is to models with irrational agents.

We consider a finite set of agents $\mathcal{I}$ partitioned into the *rational agents* $\mathcal{I}_R$ and *irrational agents* $\mathcal{I}_{IR}$. There are L goods. Each $i \in \mathcal{I}$ has a nonzero endowment of these goods $e^i > 0$ and a complete, transitive, and locally nonsatiated preference relation $\succsim^i$ whose corresponding strict preference relation is $\succ^i$.[3]

[3]Local nonsatiation means that for each $x^i \in \mathbb{R}^L_+$ and $\varepsilon > 0$ there is a $y^i \in \mathbb{R}^L_+$ such that $y^i \succ^i x^i$ and $\left\|x^i - y^i\right\| < \varepsilon$. For vector inequalities we use the notation: $x \geq y \Leftrightarrow x_k \geq y_k$ for all coordinates k,

The rational agents behave like standard consumers. Given a price vector $p \geq 0$, a rational agent i chooses a bundle from the budget set $B^i = \{x^i \in \mathbb{R}^L_+ : p \cdot x^i = p \cdot e^i\}$ that is weakly preferred to all other bundles in B^i. The irrational agents also have preference relations but can err when making constrained optimization decisions. We therefore assume only that an irrational agent i chooses some bundle from B^i.

An *exchange equilibrium* is a $(p \geq 0, x = (x^i)_{i \in \mathcal{I}})$ such that

- $\sum_{i \in \mathcal{I}} x^i \leq \sum_{i \in \mathcal{I}} e^i$,
- $x^i \in B^i$ for all $i \in \mathcal{I}$,
- if $i \in \mathcal{I}_R$ and $\widetilde{x}^i \in B^i$ then $x^i \succsim^i \widetilde{x}^i$.

The irrationality of agents need not threaten the existence of exchange equilibria, which requires the continuity of agents' demand functions but not the rationality of preferences.

We say that a coalition of agents $\mathcal{C}$ can *achieve* $(\widetilde{x}^i)_{i \in \mathcal{C}}$ *by exiting* if $\sum_{i \in \mathcal{C}} \widetilde{x}^i \leq \sum_{i \in \mathcal{C}} e^i$, and that an exchange equilibrium (p, x) is *Pareto efficient for the rational agents* if there does not exist a $(\widetilde{x}^i)_{i \in \mathcal{I}_R}$ that $\mathcal{I}_R$ can achieve by exiting such that $\widetilde{x}^i \succsim^i x^i$ for all $i \in \mathcal{I}_R$, and $\widetilde{x}^i \succ^i x^i$ for some $i \in \mathcal{I}_R$.

In the language of cooperative game theory, a coalition of agents 'blocks' an allocation x if the coalition can achieve an allocation by exiting that makes every i in the coalition at least as well off as at x^i and at least one i in the coalition better off. Thus an equilibrium (p, x) is Pareto efficient for the rational agents if the rational agents cannot block x. Feasible allocations that cannot be blocked by any coalition are in the 'core.' Since the irrational agents choose arbitrary rather than optimal bundles from their budget sets, the allocation x of an exchange equilibrium with irrational agents will usually not be Pareto optimal; hence x could be blocked by $\mathcal{I}$ and is not in the economy's core. But the same argument that shows that an arbitrary coalition of agents cannot block a standard competitive equilibrium applies to the coalition of rational agents. Thus the irrational agents, though they stand in the way of full Pareto optimality, will not lead the rational agents to split off on their own.

$x > y \Leftrightarrow (x \geq y$ and $x \neq y)$, and $x \gg y \Leftrightarrow x_k > y_k$ for all coordinates k.

Proposition 1 *Exchange equilibria are Pareto efficient for the rational agents regardless of the number of irrational agents.*

Proofs are in the appendix. Proposition 1 does not say that the coalition of rational agents cannot achieve a Pareto improvement by manipulating its market demands to change p and the commodity demands of the irrational agents. There could well be a $\widehat{p}$ and a feasible allocation $\widehat{x}$ such that (i) the irrational agents demand $(\widehat{x}^i)_{i\in\mathcal{I}_{IR}}$ at the price vector $\widehat{p}$ and (ii) the rational agents are all strictly better off with $\widehat{x}$ than at the exchange equilibrium. See McFadden (1969).

Proposition 1 does not address whether the rational agents gain anything from their trades with the irrational agents. Could the rational agents do equally well by themselves? If the irrational agents have trade with nonzero value with the rational agents and the rational agents have strictly convex preferences, there is an unambiguous answer.[4]

Proposition 2 *If in an exchange equilibrium (p, x) the rational agents have trade with nonzero value with the irrational agents ($p(k)\sum_{i\in\mathcal{I}_R}(x^i(k) - e^i(k)) \neq 0$ for some good k) and if the rational agents have strictly convex preferences, then any allocation that the rational agents can achieve by exiting leaves at least one rational agent i worse off than at i's equilibrium allocation.*

The significance of Proposition 2 lies in the contrast to games that lead to allocations that are Pareto efficient for the rational. When rational and irrational agents interact through competitive markets, irrational agents still have something to offer: they allow the rational agents to achieve welfare levels that they could not achieve on their own. We will see in section 4 that games with irrational agents need not share this property.

The nonzero trade condition in Proposition 2 is generic: if we use a standard parameterization of agents' excess demand functions then, for almost every model, in each competitive equilibrium any set of agents will have trade with nonzero value with the remaining agents. Proposition 2 can therefore be read as a remark that competitive equilibria cannot be fragmented; whether or not some agents are irrational, a competitive allocation typically cannot be achieved by partitioning the set of agents into blocs who do not trade with each other.

[4] Agent i has strictly convex preferences if $x^i \succsim^i y^i$, $x^i \neq y^i$, and $\lambda \in (0,1)$ imply $\lambda x^i + (1-\lambda)y^i \succ^i y^i$.

3 Irrational firms

Even when Pareto efficiency for the rational agents obtains, irrational consumers still cause harm in that they obstruct full Pareto optimality. Irrational producers – for example, firms that stick with a backward technology or that indulge a desire to discriminate in hiring – need not lead to any harm at all.

Let Y_j be the production set of firm j. We assume that each firm j is capable of inaction, $0 \in Y_j$. Given p, a rational firm j chooses a profit-maximizing y_j in Y_j whereas an irrational firm may take any action as long as it does not go bankrupt. So the only restriction on an irrational firm j is that it choose a $y_j \in Y_j$ such that $p \cdot y_j \geq 0$.

We label the economy's finite set of firms $\mathcal{F}$, partitioned into the rational firms $\mathcal{F}_R$ and irrational firms $\mathcal{F}_{IR}$. Define the aggregate production set of the rational firms, $Y_R = \sum_{j \in \mathcal{F}_R} Y_j$, the aggregate production set of the irrational firms, $Y_{IR} = \sum_{j \in \mathcal{F}_{IR}} Y_j$, and the aggregate production set overall, $Y = Y_R + Y_{IR} = \sum_{j \in \mathcal{F}} Y_j$.

Let $\theta_{ij} \geq 0$ be the ownership share of consumer i in firm j where, for each firm j, $\sum_{i \in \mathcal{I}} \theta_{ij} = 1$. Given p and the production decisions $(y_j)_{j \in \mathcal{F}}$, consumer i's profit income is $\sum_{j \in \mathcal{F}} \theta_{ij} p \cdot y_j$, and so the budget set for agent i is now $B^i = \{x^i \in R^L_+ : p \cdot x^i = p \cdot e^i + \sum_{j \in \mathcal{F}} \theta_{ij} p \cdot y_j\}$. A *production equilibrium* is a $(p \geq 0,\ x = (x^i)_{i \in \mathcal{I}},\ y = (y_j)_{j \in \mathcal{F}})$ such that

- $\sum_{i \in \mathcal{I}} x^i \leq \sum_{i \in \mathcal{I}} e^i + \sum_{j \in \mathcal{F}} y_j$,
- $x^i \in B^i$ for all $i \in \mathcal{I}$,
- if $i \in \mathcal{I}_R$ and $\widetilde{x}^i \in B^i$ then $x^i \succsim^i \widetilde{x}^i$,
- $y_j \in Y_j$ for all $j \in \mathcal{F}$,
- if $j \in \mathcal{F}_R$ and $\widetilde{y}_j \in Y_j$ then $p \cdot y_j \geq p \cdot \widetilde{y}_j$,
- if $j \in \mathcal{F}_{IR}$ then $p \cdot y_j \geq 0$.

Proposition 1 – that equilibria are Pareto efficient for the rational agents no matter the number of irrational agents – extends to production equilibria. The only wrinkle concerns the production possibilities that are available to the rational agents if they exit.

When a rational firm j that exits is wholly owned by rational consumers, the departing rational agents should have access to all of Y_j. But if a rational firm j is partly owned by irrational consumers, the irrational consumers who remain behind should not be denied all use of j's technology. We could let both the stayers and exiters use of all of Y_j, a legitimate solution if Y_j satisfies constant returns to scale (CRS).[5] But if Y_j shows decreasing returns to scale, then the stayers and exiters would collectively be capable of productions that the unified economy had not been able to produce. To avoid this problem, we assume that, for any rational firm j, the rational agents when they exit can use a scaled-down version of any production available to j, where the scaling factor must be less than or equal to the share of firm j that the rational consumers own: for each $j \in \mathcal{F}_R$ we set some nonnegative $\mu_j \leq \sum_{i \in \mathcal{I}_R} \theta_{ij}$ and let the rational agents when they exit use any production bundle that equals $\mu_j y_j$ for some $y_j \in Y_j$. The bound on μ_j prevents a partitioned economy from producing previously unavailable bundles. One reasonable way to proceed would be to set $\mu_j = 0$ if j is at least partly owned by irrational consumers and shows decreasing returns, $\mu_j = 1$ if j is wholly owned by rational consumers, and $\mu_j > 0$ if j is partly owned by rational consumers and satisfies CRS (which in the last case would let the rational agents use any $y_j \in Y_j$).

To extend Proposition 1, define a production equilibrium (p, x, y) to be *Pareto efficient for the rational agents* if there does not exist a $((\widetilde{x}^i)_{i \in \mathcal{I}_R}, (\widetilde{y}_j)_{j \in \mathcal{F}_R})$ such that $\widetilde{x}^i \succsim^i x^i$ for each $i \in \mathcal{I}_R$ and with strict preference for some $i \in \mathcal{I}_R$, $\widetilde{y}_j \in Y_j$ for each $j \in \mathcal{F}_R$, and

$$\sum_{i \in \mathcal{I}_R} \widetilde{x}^i \leq \sum_{i \in \mathcal{I}_R} e^i + \sum_{j \in \mathcal{F}_R} \mu_j \widetilde{y}_j.$$

Assuming that consumer preferences satisfy the assumptions of the previous section, we

[5] A production set Y_j satisfies CRS if $\lambda \geq 0$ and $y_j \in Y_j$ imply $\lambda y_j \in Y_j$. Normally CRS has no content: for any Y_j that exhibits decreasing returns we can invent a new commodity input specific to firm j, distributed to consumers to match their ownership share in j, and a new CRS production set $\overline{Y}_j$ that coincides with Y_j at the points in $\overline{Y}_j$ where the coordinate of the invented good equals 1. The behavior of the economy with these $\overline{Y}_j$ will be identical to the behavior of the original economy. We could use this trick to define the productions available to departing rational agents: this is in fact the special case in the model below where $\mu_j = \sum_{i \in \mathcal{I}_R} \theta_{ij}$ for each rational firm j. But both this special case and our general model go beyond an accounting convention: since we are considering the consequences of an actual exit, letting an exiting firm j use a scaled-down version of Y_j is a substantive assumption about how technology can be subdivided.

can conclude that production equilibria are Pareto efficient for the rational agents, regardless of the number of irrational agents (see the appendix for a proof).

The more important feature of irrational firms is that they need not interfere with full production efficiency. To achieve production efficiency, two conditions must be met. First, the rational firms must have technologies that are at least as advanced as the irrational firms: $Y_R \supset Y_{IR}$. If this condition were not satisfied – if the irrational firms can produce some bundles that the rational firms cannot – then the irrational firms could produce inefficiently and still survive market competition. The second condition is a version of constant returns to scale. In a world of decreasing-returns technologies, profit-maximizing firms can earn positive profits in equilibrium. Hence one or more of these firms could instead operate inefficiently, using its profits to subsidize its inefficient production. One simple assumption (stronger than what we will impose) that would rule out this scenario would be to suppose, in addition to $Y_R \supset Y_{IR}$, that Y_R exhibits constant returns. Constant returns is not terribly demanding; it in effect requires that all inputs are marketed commodities.[6]

The assumption that we do use is weaker and folds in the requirement that $Y_R \supset Y_{IR}$. If there exists a constant-returns production set $\widehat{Y}$ such that $Y_R \supset \widehat{Y} \supset Y_{IR}$, we say that Y_R *constant-returns dominates* Y_{IR}. The main advantage of constant-returns domination over plain constant returns arises when there are industries where no irrational firms operate: in these industries, we can allow any or all of the rational firms to exhibit decreasing returns.

A production equilibrium (p, x, y) is *production efficient* if there does not exist $(\widetilde{y}_j \in Y_j)_{j\in\mathcal{F}}$ such that $\sum_{j\in\mathcal{F}} \widetilde{y}_j > \sum_{j\in\mathcal{F}} y_j$.

Proposition 3 *If Y_R constant-returns dominates Y_{IR} then any production equilibrium with $p \gg 0$ is production efficient, regardless of the number of irrational consumers or firms.*

[6] Once again we cannot resort to the trick of rationalizing constant returns by postulating a firm-specific input for each firm with a decreasing-returns technology. The reason however is not that we need to consider any fractional rescalings of production sets, but interference with the requirement that $Y_R \supset Y_{IR}$: if each rational firm requires a firm-specific input in order to produce, then any y in Y_{IR} that actually produces some good and does not use any of the rational firms' specific inputs cannot be in Y_R. So constant returns to scale must be given its standard substantive interpretation.

No exit of rational agents is involved in Proposition 3; production efficiency obtains despite the presence of irrational agents. If, in addition to the assumptions of Proposition 3, every consumer is rational then full Pareto efficiency obtains.

The compatibility of irrational firms and full production efficiency contrasts with the more limited efficiency-for-the-rational-agents that holds on the consumer side. This divergence parallels the different punishments that competitive markets mete out to irrational producers and consumers. The consumers of market economies need not be any more rational than their counterparts in other institutional settings; their optimization errors only bring about a utility loss. But competitive markets can drive backward producers out of business, putting firms on a tighter leash.

Formally, the economies in Proposition 3 achieve production efficiency instantaneously. A more realistic picture emerges if we apply the constant returns to scale condition only to the long run; then irrational firms can survive for a while and are only driven slowly from the market. Constant returns and hence constant-returns domination are questionable when imposed on production for the near future since outputs in the near future require inputs, such as installed capital equipment, that are not marketed commodities. In these shorter time frames where decreasing returns prevails, inefficient irrational firms can survive. But constant returns to scale or constant-returns domination is plausible when imposed on production for the more distant future since all inputs should then be purchasable. If production activities for the immediate future are separable from activities for the more distant future, and irrational firms do not use their short-run profits to cross-subsidize long-run production, the logic of Proposition 3 will eventually apply: after enough time passes, production efficiency will obtain.

For an example of how the dynamic path to full efficiency plays out, let time run from 1 to T and suppose each firm j produces a single good at each date. Assume for each period t that the production set of firm j for its output at t, $Y_j(t)$, uses an input stream that lasts for τ periods. The outputs that appear before date τ therefore use inputs prior to date 1 but these inputs are not included among the economy's L goods.[7] Let the rational firms have technology that is at least as good as the technology of the irrational

[7] As always, each $Y_j(t)$ is a subset of $\mathbb{R}^L$.

firms: $\sum_{j\in\mathcal{F}_R}\sum_{t=1}^{T} Y_j(t) \supset \sum_{j\in\mathcal{F}_{IR}}\sum_{t=1}^{T} Y_j(t)$. Because the production of outputs at early dates requires inputs that have to be applied before date 1, the $Y_j(t)$ for $t < \tau$ may exhibit decreasing returns to scale, but suppose that the $Y_j(t)$ exhibit constant returns to scale for $t \geq \tau$. Then, as t increases, the number of outputs produced under constant returns to scale increases. The economy can proceed through time by letting a firm j incur debt when it begins the purchase of an input stream and then paying off this debt and distributing any profits to its shareholders when the output appears τ periods later, thus ruling out cross-subsidization. In an equilibrium where the rational firms maximize profits and all firms must earn nonnegative profits, Proposition 3 applies to all the outputs produced under constant returns. The number of outputs whose production is efficient therefore steadily increases through time. If an inflow of new firms introduces more advanced technology into some existing sectors and if these entrants do not maximize profits, then production inefficiency could obtain in some sectors of the economy while it is being driven out in the innovation-free sectors. Competitive general equilibrium models with irrational agents can thus give a Schumpeterian account of firm entry-exit dynamics.

In a competitive equilibrium model that contains only rational agents, a firm with sufficiently backward technology will shut itself down. In the present model, as in the Schumpeterian tradition, irrational firms with backward technology firms must be driven out of business.[8] Although the two mechanisms will often cause the same firms to exit they can lead to sharply different policy advice. Consider, for instance, the traditional design of trade liberalization and deregulation policies that harm firms that have been protected from market competition. When firms or consumers could be harmed by (and might therefore obstruct) reforms, the classical welfare theorems show how to engineer Pareto improvements using lump-sum compensation payments. But with irrational firms, lump-sum payments can undermine production efficiency: they give irrational firms the leeway to take inefficient actions without going bankrupt. Irrational firms therefore present policy-makers with a trade-off: Pareto improvements are possible if compensation payments keep irrational firms afloat but then production efficiency will be undermined.

[8] See Klette and Kortum (2004) and Lentz and Mortensen (2005) for modern Schumpeterian approaches.

Either production efficiency or a Pareto improvement can be achieved but not both.[9]

Despite the common ground with Schumpeter, the logic presented here for why the rational and efficient firms come to predominate differs from the evolutionary mechanisms that the main Schumpeterian modeling tradition has relied on. In Proposition 3, production efficiency is achieved entirely through the price system, as the rational firms drive the irrational firms from the market. In evolutionary models (e.g., Nelson and Winter (1982) and earlier Alchian (1950) and Friedman (1953)), in contrast, the efficient firms become more prevalent because they make larger profits and grow faster.

To summarize, the theory of production efficiency in this paper is distinct from both the evolutionary and Arrow-Debreu explanations. The present account operates via the price mechanism but does without the Arrow-Debreu assumption of universal rationality.

4 Games with irrational agents

The irrationality-proofness of efficiency in competitive markets raises the question of whether markets are distinctive in this regard. Can games do as well as markets? To compare like with like, we consider games that fully implement competitive allocations and assess the irrationality-proofness of their efficiency properties. Under the assumption that all agents are rational, the starting point of Walrasian and Nash equilibria, markets and games that implement competitive outcomes bear a close resemblance. But differences come out when we let some agents be irrational. The typical constructions of classical implementation theory are brittle: just one irrational agent can lead to nonexistence of equilibria or inefficiency for the rational agents. But better-performing games can be designed.

There are many ways to define equilibrium in games with irrational agents. We take the view, discussed in the introduction, that irrational agents can be predictable, and therefore define equilibrium as a strategy profile such that the strategy of each rational agent is a best response to the strategies of all other agents, whether they be rational or irrational, just as in a Nash equilibrium of a standard game every agent best responds

[9] The working paper version of this article uses the example of trade liberalization to illustrate this dilemma.

to the strategies played by the other agents. Our definition of equilibrium thus stakes out a compromise: we let some agents be irrational, avoiding the full-bore rationality assumptions of Nash implementation, but also let rational agents best-respond to the actions of irrational agents – the rationals do not think the irrationals as so erratic that they must play strategies that are optimal no matter how the irrationals act. The value of the compromise is that desirable equilibria will exist. As we will see, a dominant-strategy approach would lead to an impossibility result.

To keep the parallels between markets and games tight, we consider mechanisms that fully implement the outcomes that are targeted: when all agents are rational, the set of equilibrium outcomes and the set of competitive allocations will coincide exactly, just as all competitive equilibria lead to competitive allocations when all agents are rational.

There are again L goods, and each agent $i \in \mathcal{I}$ has an endowment $e^i \gg 0$ and preferences $\succsim^i$, defined over nonnegative bundles of the L goods, that are complete, transitive, monotone, continuous, and convex.[10] We fix the endowment profile $(e^1, ..., e^I)$ throughout, where I is the number of agents.[11] If there are competitive allocations on the boundary of agents' consumptions sets then those allocations would not be Nash implementable (Hurwicz, Maskin, and Postlewaite (HMP) (1995)). So, even in the absence of irrational agents, there would be no game whose equilibria exactly coincide with the competitive allocations, hampering comparison to the rest of the paper. To step around this problem, we make an *interiority* assumption that each i's indifference curve through e^i does not intersect the coordinate axes.[12] When agents' preferences $(\succsim^i)_{i\in\mathcal{I}}$ satisfy the assumptions of this paragraph, we say that $(\succsim^i)_{i\in\mathcal{I}}$ (or simply the *model*) is *admissible*. For any admissible model, an exchange equilibrium (as defined in section 2) exists. Let e denote $\sum_{i\in\mathcal{I}} e^i$.

A *mechanism* is defined by strategy sets S^i for $i \in \mathcal{I}$ and an outcome function g that maps each strategy profile $s = (s^1, ..., s^I)$ to a feasible allocation $x = (x^1, ..., x^I)$, that is,

[10] The preferences $\succsim^i$ are *monotone* if $x^i > z^i$ implies $x^i \succ^i z^i$, *convex* if $\lambda \in [0, 1]$ and $x^i \succsim^i z^i$ imply $\lambda x^i + (1-\lambda)z^i \succsim^i z^i$, and *continuous* if $\{x^i \in \mathbb{R}^L_+ : x^i \succsim^i z^i\}$ and $\{x^i \in \mathbb{R}^L_+ : z^i \succsim^i x^i\}$ are closed sets for all $z^i \in \mathbb{R}^L_+$.

[11] Equivalently, we could let $e = (e^1, ..., e^I)$ vary and assume that the game designer knows e and can use this information along with agents' strategy choices to determine allocations. We could also let e be revealed by agents' strategy choices.

[12] Formally, $\succsim^i$ satisfies interiority if, for all $x^i \in \mathbb{R}^L_+$, $x^i \succsim^i e^i \Rightarrow x^i \gg 0$.

a x where $\sum_{i\in\mathcal{I}} x^i \leq e$. Given a mechanism and an admissible $(\succsim^i)_{i\in\mathcal{I}}$, an *equilibrium* is a pair $(s = (s^1, ..., s^I), \mathcal{I}_R)$ that specifies the strategies that all agents play and a set of rational agents $\mathcal{I}_R$ such that, for each $i \in \mathcal{I}_R$,

$$g^i(s) \succsim^i g^i(\widetilde{s}^i, s^{-i}) \text{ for all } \widetilde{s}^i \in S^i.$$

So each rational i in an equilibrium $(s, \mathcal{I}_R)$ is optimizing with s^i given that the other agents play s^{-i}.

A mechanism $((S^i)_{i\in\mathcal{I}}, g)$ *implements competitive allocations when the set of rational agents is* $\mathcal{I}_R$ if, for any admissible model $(\succsim^i)_{i\in\mathcal{I}}$ and any allocation x,

$$\begin{aligned}&\text{(there is an equilibrium } (s, \mathcal{I}_R) \text{ such that } x = g(s)) \Leftrightarrow \\ &\qquad \big((p, x) \text{ is an exchange equilibrium for } (\succsim^i)_{i\in\mathcal{I}} \text{ for some } p\big).\end{aligned}$$

In words, a mechanism implements competitive allocations when $\mathcal{I}_R$ is the set of rational agents if, for all preference profiles, any equilibrium outcome of the mechanism when $\mathcal{I}_R$ is the set of rational agents is a competitive allocation, and conversely for any competitive allocation there is an equilibrium of the mechanism when $\mathcal{I}_R$ is the set of rational agents whose outcome is that allocation. Although it will turn out that competitive allocations cannot be implemented when some agents are irrational – just as they cannot be in an exchange economy with irrational agents – our definition is designed to say what the implementation of competitive allocations would mean in such cases.

The above definition indicates our use of a 'full' concept of implementation: the sets of equilibrium outcomes and competitive allocations coincide when all agents are rational. A weaker definition would require only that for each competitive allocation x there is some equilibrium that reaches x but would allow other equilibria to reach noncompetitive allocations. The fit with the competitive markets would then be looser, since competitive equilibria always generate competitive allocations when all agents are rational. But in addition the weaker definition of implementation would be too permissive. Some games that for each competitive allocation x have an equilibrium that reaches x can also

implement a vast set of other allocations.[13]

We first consider a specific mechanism, a much simplified version of a mechanism in HMP, that fully implements competitive allocations when all agents are rational. The introduction of irrational agents into this game blocks even the existence of equilibrium.

Example 1 Each $S^i = \{(p^i, x^i) \in (\mathbb{R}^L_+ \backslash \{0\}) \times \mathbb{R}^L_+ : p^i \cdot x^i = p^i \cdot e^i\}$. Given the strategies $(p^i, x^i)_{i\in\mathcal{I}}$, let $P = \{p : p = p^i \text{ for some } i\}$ denote the set of price vectors that the agents announce, and let $\#P$ be the number of distinct announced price vectors. The outcome $g((p^i, x^i)_{i\in\mathcal{I}})$ of the mechanism is then the allocation $(\overline{x}^i)_{i\in\mathcal{I}}$ defined by:

1: if $\#P = 1$ and $\sum_{i\in\mathcal{I}} x^i = e$, then

$$\overline{x}^i = x^i \text{ for all } i \in \mathcal{I},$$

2: if $\#P = 2$ and there is a k such that $\left\|p^k\right\| > \|p^i\|$ for $i \neq k$, $x^k \leq e$ and $p^i \cdot x^k = p^i \cdot e^k$, then

$$\overline{x}^k = x^k \text{ and } \overline{x}^i = 0 \text{ for } i \neq k,$$

3: if $\#P > 2$ and there is a k such that $\left\|p^k\right\| > \|p^i\|$ for $i \neq k$, then

$$\overline{x}^k = e \text{ and } \overline{x}^i = 0 \text{ for } i \neq k,$$

4: in all other cases, $\overline{x}^i = 0$ for all $i \in \mathcal{I}$.

It is easy to confirm that the allocation of any exchange equilibrium $(p, (x^i)_{i\in\mathcal{I}})$ is an equilibrium outcome of this mechanism when all agents are rational, $\mathcal{I}_R = \mathcal{I}$. In the equilibrium, each agent i names p and x^i: if there are at least two agents, rules 2 and 4 imply that any unilateral deviation for an agent k can at best lead to a $\widetilde{x}^k$ such that $p \cdot \widetilde{x}^k = p \cdot e^k$. Conversely, given an equilibrium of the mechanism, $(p^i, x^i)_{i\in\mathcal{I}}$, rules 2, 3, and 4 imply that if two or more agents name different prices then only an agent k who names a price vector such that $\left\|p^k\right\| > \|p^i\|$ for all $i \neq k$ will avoid the 0 bundle; since there can be only one such k, the equilibrium must have a unanimous announcement of prices p. And the x^i must satisfy $\sum_{i\in\mathcal{I}} x^i = e$ since otherwise any agent would take advantage of rule 2 to avoid the 0 bundle. Rule 2 also implies that any agent k could by deviating achieve any $\widetilde{x}^k \leq e$ such that $p \cdot \widetilde{x}^k = p \cdot e^k$. Given that e^k is therefore achievable, the interiority assumption implies that each $x^k \gg 0$. Since $\sum_{i\in\mathcal{I}} x^i = e$, we have $e \gg x^i$ for

[13] Consider, for example, a mechanism where all agents name an allocation x; if everyone names the same x then each i receives x^i and otherwise everyone receives the 0 bundle. Any allocation is then an equilibrium outcome when all agents are rational.

all i, assuming that there are at least two agents. It then follows from convexity that for any i there is no $\widetilde{x}^i \succ^i x^i$ with $p \cdot \widetilde{x}^i = p \cdot e^i$, whether $\widetilde{x}^i$ is feasible or not (see the proof of Proposition 4 for more detail on this point). Thus the equilibrium delivers a competitive allocation.[14]

The all-rational equilibria are delicate however. If the population includes irrational agents who choose arbitrary elements of S^k then typically there will be no equilibria. If two or more irrational agents choose different prices and there are two or more rational agents, then there will be no profile of optimizing strategies for the rational agents (just as in the previous paragraph). If there is a single irrational agent i, still with two or more rational agents, then unless x^i and the Walrasian demands of the rational agents at p^i happen to sum to e there will be no equilibrium: there again could not be a unanimous announcement of prices since one of the rational agents would take advantage of rule 2 to name a different price and achieve his Walrasian demand. Existence of equilibrium thus displays a minimal level of irrationality-proofness: if there are two rational agents then just one irrational agent is enough to prevent there from being an equilibrium. ■

The irrationality proofness problem of the above example is that equilibria fail to exist when irrational agents are present. Other mechanisms that implement competitive allocations when all agents are rational, e.g., Jackson et al. (1994) which uses undominated Nash equilibria, always have equilibria when irrational agents are present but fail to achieve Pareto efficiency for the rational agents.

Are there mechanisms that, when all agents are rational, fully implement competitive allocations and, when some agents are irrational, not only have equilibria but have only equilibria that achieve Pareto efficiency for the rational agents?

As in section 2, a coalition $\mathcal{C}$ can achieve $(x^i)_{i\in\mathcal{C}}$ by exiting if $\sum_{i\in\mathcal{C}} x^i \leq \sum_{i\in\mathcal{C}} e^i$. Given $\mathcal{I}_R$, let us say that the allocation $(x^i)_{i\in\mathcal{I}}$ is *Pareto efficient for the rational agents* if there is no $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$ that $\mathcal{I}_R$ can achieve by exiting such that $\widetilde{x}^i \succsim^i x^i$ for all $i \in \mathcal{I}_R$ and $\widetilde{x}^i \succ^i x^i$ for some $i \in \mathcal{I}_R$. Finally, a mechanism $((S^i)_{i\in\mathcal{I}}, g)$ is *Pareto efficient for the*

[14] If there is a single agent then rules 1 and 4 imply that the only equilibrium allocation of the mechanism is e^1, which is the competitive allocation. The fact that the mechanism implements competitive allocations when all agents are rational without a restriction on the number of agents is due to our assumption that goods can be freely disposed of.

rational agents if, for any admissible $(\succsim^i)_{i\in\mathcal{I}}$, any set of rational agents $\mathcal{I}_R \subset \mathcal{I}$, and any strategy profile for the irrational agents $(s^i)_{i\in\mathcal{I}_{IR}}$, there is an equilibrium $(s, \mathcal{I}_R)$ where the irrational agents play $(s^i)_{i\in\mathcal{I}_{IR}}$ and, for every equilibrium $(s, \mathcal{I}_R)$ when the irrationals play $(s^i)_{i\in\mathcal{I}_{IR}}$, $g(s)$ is Pareto efficient for the rational agents. In line with our definition of the implementation of competitive allocations, we require that every equilibrium, not just one, leads to an outcome that is Pareto efficient for the rational.

Proposition 4 *There are mechanisms that are both Pareto efficient for the rational agents and implement competitive allocations when all agents are rational.*

Thus there are games that perform reasonably well regardless of the number of irrational agents and how they play. The proof of Proposition 4 designs a mechanism with two stages of competition. The first stage ensures that if some or all of the irrational agents choose strategies that are inconsistent with an outcome that is Pareto efficient for the rational then the rational agents can split off on their own; they can defeat some or all of the irrational agents in an integer game and determine a final allocation using only their own resources and the resources of any irrational agents who do happen to choose compatible strategies. A 'victorious bloc' that contains all of the rational agents thus emerges from the first stage. The second stage is more traditional and is similar to the HMP mechanism. If a single agent in the victorious bloc deviates from a candidate equilibrium the deviator can achieve only those bundles that are in a budget set defined by prices that the agents in the victorious bloc simultaneously announce. Multiple deviations on the other hand set off an unwinnable integer game where everyone but the winner receives an undesirable bundle. As usual this device blocks outcomes that the mechanism aims to avoid (in our case, the allocations that fail to be Pareto efficient for the rational). The integer games in the two stages thus serve opposite purposes: in the first, some or all of the irrationals may well be defeated in equilibrium while in the second there can be no winner.

The universe of possible mechanisms displays such strategic variety that one might wonder if we can do better: are there mechanisms that implement competitive allocations even when irrational agents are present? The answer is 'no.' For suppose two models, 1 and 2, differ only in the preferences of the irrational agents and let each have a unique

competitive allocation that differs from the competitive allocation of the other model. If competitive allocations were always achieved then, in either model k, for any strategy profile that the irrational agents might play, there would be a profile of equilibrium strategies for the rational agents that leads to the competitive outcome of model k. But if, say, model 1 obtains and the irrational agents play some profile $(s^i)_{i\in\mathcal{I}_{IR}}$, it will be an equilibrium for the rational agents to play the profile that they play in model 2 when the irrational play $(s^i)_{i\in\mathcal{I}_{IR}}$ (since the rational agents' preferences are unchanged).

Proposition 5 *There is no mechanism that implements competitive allocations when* $\mathcal{I}_R \neq \mathcal{I}$ *(i.e., some agent is irrational).*

Games and competitive markets therefore share some common ground. As with markets, games with irrational agents cannot always reach a competitive outcome but they can achieve Pareto efficiency for the rational agents. Still there is an important difference between markets and games. In the mechanisms that underlie Proposition 4, when irrational agents do not choose strategies compatible with an efficient allocation the rational agents trump them and split off on their own. In fact, the rational agents will typically end up with bundles that in the aggregate use only their own endowments. So in games the property of 'Pareto efficiency for the rational agents' does not imply that the rational agents will gain from the presence of irrational agents. In contrast, as we saw in section 2, in a competitive equilibrium the rational agents will generically trade with the irrational agents and thus achieve welfare levels that they could not achieve on their own (see Proposition 2). In this generic sense, markets can outperform full-implementation games: they automatically use the resources of the irrational agents to make the rationals better off.

If the strategic actions of irrational agents cannot be predicted, then our definition of equilibrium is open to criticism. In our equilibria, each rational agent best responds to the strategies that all other agents play, whether they are rational or irrational; implicitly, the rational agents know how the irrational play. To accommodate unpredictable irrational agents, we could require that each rational agent's strategy is a best response to the other agents' strategies, whatever set of agents turns out to be irrational and for all strategy profiles that the irrational agents might play. Since any agent can be irrational, this would

require that rational agents play weakly dominant strategies. But then unfortunately there cannot be a mechanism that is Pareto efficient for the rational agents. For suppose to the contrary that there were such a mechanism. Then, when all agents are rational and play their dominant strategies, a core allocation would have to result: if instead an allocation x were to occur that some coalition $\mathcal{C}$ could block, then when $\mathcal{C}$ is the set of rational agents and the irrational agents happen to play their dominant strategies, x would ensue and x cannot be Pareto efficient for the rational (since $\mathcal{I}_R = \mathcal{C}$ and $\mathcal{C}$ can block x). Since there is no dominant-strategy mechanism in an exchange economy setting whose outcomes consist only of core allocations, we conclude that there is no mechanism that is Pareto efficient for the rational agents when rational agents are required to play dominant strategies.[15]

Eliaz (2002), an innovative theory of implementation that allows for irrational agents, takes a different tack and requires each rational agent to play a strategy that is optimal no matter who is irrational and how they move. The Eliaz model avoids the roadblock that accompanies dominant-strategy implementation by restricting the number of irrational agents. In contrast, we placed no restrictions on the number of irrational agents in this section or in section 2 and only minimal implicit restrictions in section 3.

We have avoided any hint of Bayesian implementation; all of our agents implicitly have the same information. Had we permitted asymmetric information, there would have been no hope for Pareto efficiency for the rational agents. Implementation of efficient outcomes in the face of asymmetric information would require players with knowledge of other agents' characteristics to patrol those individuals, e.g., report their characteristics to prevent them from misrepresenting themselves. Since irrational agents might fail to undertake patrolling strategies, they can convert a model with nonexclusive information (no single agent has privileged information) into a model with exclusive information. In incomplete information settings, therefore, a single irrational agent can dramatically alter what can be implemented.[16]

[15] See Serizawa (2002) for stronger impossibility results that imply that there is no mechanism that implements only core allocations in our setting. Earlier results of this nature reach back to the seminal Hurwicz (1972) and include Dasgupta et al. (1979), Satterthwaite and Sonnenschein (1981), and Zhou (1991).

[16] See Postlewaite and Schmeidler (1986) and Blume and Easley (1990) for the implementation conse-

5 Conclusion

One goal of irrationality-proofness is to serve as a robustness check. For the property of a model to be reliable, it should survive the introduction of irrational agents who do not trade or choose strategies optimally. By recasting efficiency – moving from classical Pareto efficiency to production efficiency and to Pareto efficiency for the rational agents – competitive equilibria can pass the robustness test. Indeed, these alternative definitions of efficiency can withstand the introduction of large numbers of irrational agents. Our conclusions are driven by the separating feature of prices: if rational consumers and firms face a common price vector then constrained forms of efficiency will hold, even when irrational agents are present.

Our analysis of games shows that no inevitable divide between the irrationality-proofness of efficiency in games and in competitive markets. But in contrast to markets, the conclusion that irrational agents in a game do little harm requires a careful construction: we have to let the rational agents' strategies vary as a function of the irrational agents' strategies and rule out asymmetric information. From the broader perspective freed from these restriction, the irrationality-proofness of efficiency is more robust for markets than for games.

6 Appendix: Proofs

Proof of Proposition 1. Let (p, x) be the competitive equilibrium. If the rational agents can achieve $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$ by exiting, then $\sum_{i\in\mathcal{I}_R}\widetilde{x}^i \leq \sum_{i\in\mathcal{I}_R} e^i$. Multiply by p to get (1) $p\cdot\sum_{i\in\mathcal{I}_R}\widetilde{x}^i \leq p\cdot\sum_{i\in\mathcal{I}_R} e^i$. If $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$ is a Pareto improvement for the rational agents, then (2) $\widetilde{x}^i \succsim^i x^i$ for all $i\in\mathcal{I}_R$, and (3) $\widetilde{x}^h \succ^h x^h$ for some $h\in\mathcal{I}_R$. Given the optimization of the rational agents, (3) implies $p\cdot\widetilde{x}^h > p\cdot e^h$, and, since each $\succsim^i$ is transitive and locally nonsatiated for each $i\in\mathcal{I}_R$, (2) implies $p\cdot\widetilde{x}^i \geq p\cdot e^i$ for all $i\in\mathcal{I}_R$. Sum over $i\in\mathcal{I}_R$ to get $p\cdot\sum_{i\in\mathcal{I}_R}\widetilde{x}^i > p\cdot\sum_{i\in\mathcal{I}_R} e^i$, contradicting (1). ■

Proof of Proposition 2. First observe that if (p, x) is the exchange equilibrium and $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$

quences of exclusive information.

satisfies

$$\sum_{i\in\mathcal{I}_R}\widetilde{x}^i \leq \sum_{i\in\mathcal{I}_R} e^i + \varphi \sum_{i\in\mathcal{I}_{IR}} (e^i - x^i) \qquad \text{(i)}$$

for some $\varphi \in \mathbb{R}$ then (1) in the proof of Proposition 1 will obtain (since $p\cdot\sum_{i\in\mathcal{I}_{IR}}(e^i-x^i) = 0$). Hence if $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$ satisfies (i) then $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$ does not Pareto dominate $(x^i)_{i\in\mathcal{I}_R}$.

If $\mathcal{I}_R$ can achieve $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$ by exiting and $\widetilde{x}^i \succsim^i x^i$ for all $i \in \mathcal{I}_R$, then Proposition 1 implies $\widetilde{x}^i \sim^i x^i$ for all $i \in \mathcal{I}_R$. Since there is a k with $p(k)\sum_{i\in\mathcal{I}_R}(x^i(k) - e^i(k)) \neq 0$, and since $p\cdot\sum_{i\in\mathcal{I}_R}(x^i-e^i) = 0$, there must be a l with $p(l) > 0$ such that $\sum_{i\in\mathcal{I}_R}(x^i(l)-e^i(l)) > 0$. Since any allocation $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$ achieved by exiting must satisfy $\sum_{i\in\mathcal{I}_R}\widetilde{x}^i \leq \sum_{i\in\mathcal{I}_R} e^i$, we conclude that $(\widetilde{x}^i)_{i\in\mathcal{I}_R} \neq (x^i)_{i\in\mathcal{I}_R}$ and therefore $\widetilde{x}^h \neq x^h$ for some $h \in \mathcal{I}_R$. By strict convexity, if $\lambda \in (0,1)$ then $\lambda x^h + (1-\lambda)\widetilde{x}^h \succ^h x^h$. Thus $(\lambda x^i + (1-\lambda)\widetilde{x}^i)_{i\in\mathcal{I}_R}$ Pareto dominates $(x^i)_{i\in\mathcal{I}_R}$. Since however,

$$\begin{aligned}\sum_{i\in\mathcal{I}_R} x^i &\leq \sum_{i\in\mathcal{I}_R} e^i + \sum_{i\in\mathcal{I}_{IR}} (e^i - x^i), \text{ and}\\ \sum_{i\in\mathcal{I}_R}\widetilde{x}^i &\leq \sum_{i\in\mathcal{I}_R} e^i,\end{aligned}$$

we have

$$\sum_{i\in\mathcal{I}_R}(\lambda x^i + (1-\lambda)\widetilde{x}^i) \leq \sum_{i\in\mathcal{I}_R} e^i + \lambda\sum_{i\in\mathcal{I}_{IR}}(e^i - x^i).$$

Since $(\lambda x^i + (1-\lambda)\widetilde{x}^i)_{i\in\mathcal{I}_R}$ therefore satisfies (i), we have a contradiction. ■

Proof of Extension of Proposition 1. Only a couple changes to the proof of Proposition 1 are needed. If the rational agents can achieve a Pareto improvement by exiting, there exist $(\widetilde{x}^i)_{i\in\mathcal{I}_R}$ and $(\widetilde{y}_j)_{j\in\mathcal{F}_R}$, where each $\widetilde{y}_j \in Y_j$, such that $\widetilde{x}^i \succsim^i x^i$ for each $i \in \mathcal{I}_R$, with strict preference for some $i \in \mathcal{I}_R$, and $\sum_{i\in\mathcal{I}_R}\widetilde{x}^i \leq \sum_{i\in\mathcal{I}_R} e^i + \sum_{j\in\mathcal{F}_R}\mu_j\widetilde{y}_j$. Since $\widetilde{y}_j \in Y_j$, profit maximization gives $\mu_j p\cdot y_j \geq \mu_j p\cdot\widetilde{y}_j$ for each $j \in \mathcal{F}_R$. Hence $p\cdot\sum_{i\in\mathcal{I}_R}\widetilde{x}^i \leq p\cdot\sum_{i\in\mathcal{I}_R} e^i + p\cdot\sum_{j\in\mathcal{F}_R}\mu_j y_j$. But optimization for the rational agents implies $p\cdot\widetilde{x}^i \geq p\cdot e^i + p\cdot\sum_{j\in\mathcal{F}_R}\theta_{ij}y_j$ for all $i \in \mathcal{I}_R$, with strict inequality holding for some $i \in \mathcal{I}_R$. Summing over the rational consumers and using the fact that $\mu_j \leq \sum_{i\in\mathcal{I}_R}\theta_{ij}$ gives the contradiction $p\cdot\sum_{i\in\mathcal{I}_R}\widetilde{x}^i > p\cdot\sum_{i\in\mathcal{I}_R} e^i + p\cdot\sum_{j\in\mathcal{F}_R}\mu_j y_j$. ■

Proof of Proposition 3. Let (p,x,y) be an equilibrium with $p \gg 0$ and suppose it is

not production efficient. There would then exist $(y'_j)_{j\in\mathcal{F}}$ such that $\sum_{j\in\mathcal{F}} y'_j > \sum_{j\in\mathcal{F}} y_j$. Since $p \gg 0$, $p \cdot \sum_{j\in\mathcal{F}} y'_j > p \cdot \sum_{j\in\mathcal{F}} y_j$. But since the rational firms are maximizing, $p \cdot \sum_{j\in\mathcal{F}_R} y'_j \leq p \cdot \sum_{j\in\mathcal{F}_R} y_j$. Hence $p \cdot \sum_{j\in\mathcal{F}_{IR}} y'_j > p \cdot \sum_{j\in\mathcal{F}_{IR}} y_j$ and so $p \cdot \sum_{j\in\mathcal{F}_{IR}} y'_j > 0$. Since Y_R constant-returns dominates Y_{IR} there exists a constant-returns production set $\widehat{Y}$ such that $Y_R \supset \widehat{Y} \supset Y_{IR}$. Hence there is a $\widehat{y} \in \widehat{Y}$ and $(\widehat{y}_j)_{j\in\mathcal{F}_R}$ such that $\sum_{j\in\mathcal{F}_R} \widehat{y}_j = \widehat{y} = \sum_{j\in\mathcal{F}_{IR}} y'_j$. So $p \cdot \sum_{j\in\mathcal{F}_R} \widehat{y}_j > 0$, and since $\widehat{Y}$ satisfies constant returns, for any $\alpha > 0$, $\alpha \sum_{j\in\mathcal{F}_R} \widehat{y}_j \in Y_R$. Hence for any $\alpha > 0$ there exists a $(\widetilde{y}_j)_{j\in\mathcal{F}_R}$, with each $\widetilde{y}_j \in Y_j$, such that $p \cdot \sum_{j\in\mathcal{F}_R} \widetilde{y}_j = p \cdot \alpha \sum_{j\in\mathcal{F}_R} \widehat{y}_j = \alpha \left(p \cdot \sum_{j\in\mathcal{F}_R} \widehat{y}_j \right)$, and so there must be a $j \in \mathcal{F}_R$ such that $p \cdot \widetilde{y}_j \geq \frac{1}{|\mathcal{F}_R|} \alpha \left(p \cdot \sum_{j\in\mathcal{F}_R} \widehat{y}_j \right)$. Since (i) for each $\alpha > 0$ there is a $j \in \mathcal{F}_R$ and $\widetilde{y}_j \in Y_j$ satisfying this inequality, (ii) $\frac{1}{|\mathcal{F}_R|} \left(p \cdot \sum_{j\in\mathcal{F}_R} \widehat{y}_j \right) > 0$, and (iii) there are finitely many firms, there must be at least one firm in $\mathcal{F}_R$ that can make unboundedly great profits, contradicting the assumption that $(y_j)_{j\in\mathcal{F}_R}$ are equilibrium production decisions. ■

Proof of Proposition 4. We fix the admissible model throughout. The mechanism consists of two parts. The first part determines if there is a 'victorious coalition'. The first four coordinates of a strategy s^i for agent i are relevant to this part: these are $\mathcal{C}^i \subset \mathcal{I}$ which gives i's proposal of a coalition, a 'coalition integer' $n^i \in \mathbb{N}$, a price $p^i \in \mathbb{R}^L_+ \backslash \{0\}$, and a consumption bundle $x^i \in \mathbb{R}^L_+$. Given $(\mathcal{C}^i, n^i, p^i, x^i)_{i\in\mathcal{I}}$, $\mathcal{C}$ is *victorious* iff there exists (n, p) such that (1) for each $i \in \mathcal{C}$, $\mathcal{C}^i = \mathcal{C}$, $p^i = p$, $n^i = n$, and $p \cdot x^i = p \cdot e^i$, (2) $n > n^k$ for each $k \notin \mathcal{C}$, and (3) $\sum_{i\in\mathcal{C}} x^i = \sum_{i\in\mathcal{C}} e^i$. So the agents in a victorious $\mathcal{C}$ must all propose $\mathcal{C}$, play a common n that defeats all outsiders in an integer game, announce a common price, and announce consumption bundles that are individually affordable and jointly feasible using the resources of $\mathcal{C}$. If there is no victorious coalition, the mechanism g assigns each $i \in \mathcal{I}$ the consumption bundle 0.

In the second part of the mechanism, which is relevant only if there is a victorious coalition $\mathcal{C}$, any agent in $\mathcal{C}$ can reject the bundle assigned to him in the first part. The second part of each s^i has three components: a a or r, which indicates whether i accepts or rejects i's assigned bundle, an integer m^i that determines the 'dominant' rejection, and the consumption w^i that i proposes to receive if i's rejection is dominant. Joining together the two parts of a strategy, we have, for each $i \in \mathcal{I}$, $S^i = (2^I \times \mathbb{N} \times \mathbb{R}^L_+ \backslash \{0\} \times \mathbb{R}^L_+)$ $\times$ $(\{a, r\} \times \mathbb{R}^L_+ \times \mathbb{N})$ with typical element $s^i = (\mathcal{C}^i, n^i, p^i, x^i, a \text{ or } r, w^i, m^i)$.

If $\mathcal{C}$ is victorious and, for all $i \in \mathcal{C}$, s^i announces a then we say $\mathcal{C}$ is 'unanimous'. If there is a victorious and unanimous coalition $\mathcal{C}$, then the outcome given by g is for each $i \in \mathcal{C}$ to receive the x^i given by s^i and for each $i \notin \mathcal{C}$ to receive the 0 bundle. If there is a coalition $\mathcal{C}$ that is victorious but not unanimous, define the set of rejectors $R_{\mathcal{C}} = \{i \in \mathcal{C}\colon\ i \text{ announces } r\}$. If $\#R_{\mathcal{C}} \geq 2$ the integer game in the second part of the mechanism determines the dominant rejection: if there is a $i \in \mathcal{C}$ such that $m^i > m^k$ for $k \in \mathcal{C}\backslash\{i\}$ and $w^i \leq e$, then the outcome is for i to receive w^i and each $k \in \mathcal{I}\backslash\{i\}$ to receive 0. In all other cases with $\#R_{\mathcal{C}} \geq 2$, the outcome is for each i to receive 0. If $R_{\mathcal{C}} = \{i\}$ but $p \cdot w^i \neq p \cdot e^i$ or $w^i > \sum_{k \in \mathcal{C}} e^k$ (where p is the common price announcement of the members of $\mathcal{C}$), then each $k \in \mathcal{I}$ receives 0. Finally we impose the following 'single deviation rule': if $R_{\mathcal{C}} = \{i\}$, $p \cdot w^i = p \cdot e^i$, and $w^i \leq \sum_{k \in \mathcal{C}} e^k$, then the outcome is that i receives w^i and each $k \in \mathcal{I}\backslash\{i\}$ receives 0.

We fix the strategies of $i \in \mathcal{I}_{IR}$, and let n_{IR} denote $\max\{n^i : i \in \mathcal{I}_{IR}\}$. Let $(p, x^i)_{i \in \mathcal{I}_R}$ be an exchange equilibrium for the society consisting solely of $\mathcal{I}_R$. Then $(\mathcal{I}_R, n_{IR}+1, p, x^i;$ $a, 0, 1)$ is an equilibrium. For suppose some $i \in \mathcal{I}_R$ deviates by announcing a different coalition, a different price, or a different coalition integer. If this deviation does not permit there to be a victorious coalition then i would receive 0 and so the deviation would not be undertaken. And the deviation can permit there to be a victorious coalition only if $n^i > n_{IR} + 1$ and i's coalition announcement is $\{i\}$, in which case i receives either the consumption $\widetilde{x}^i = e^i$ or $\widetilde{x}^i = 0$; since in either case, $x^i \succsim^i \widetilde{x}^i$, we again conclude that it is optimizing for i not to deviate. If on the other hand i deviates with $(\mathcal{I}_R, n_{IR} + 1, p, x^i;$ $r, w^i, m^i)$ then i receives either 0 (if $p \cdot w^i \neq p \cdot e^i$ or $w^i > \sum_{k \in \mathcal{I}_R} e^k$) or, given the definition of an exchange equilibrium, a w^i with $x^i \succsim^i w^i$. Thus for any strategy profile for the irrational agents, there is an equilibrium where the irrational agents play that profile. Furthermore, given Proposition 1 and the fact that $(p, x^i)_{i \in \mathcal{I}_R}$ is an exchange equilibrium for $\mathcal{I}_R$, the equilibrium outcome is Pareto efficient for the rational agents. In the case where $\mathcal{I}_{IR} = \varnothing$, for any exchange equilibrium (p, x), the outcome of the above equilibrium is the competitive allocation x.

It remains to show that any equilibrium outcome is Pareto efficient for the rational agents and is the competitive allocation when $\mathcal{I}_{IR} = \varnothing$. Let x be an arbitrary equilibrium

outcome. Since Pareto efficiency for the rational agents holds vacuously if $\mathcal{I}_R = \varnothing$, we suppose that $\mathcal{I}_R \neq \varnothing$. Since any agent i can receive e^i by announcing $(\{i\}, n^i, p^i, e^i; a, 0, 1)$, where $n^i > n^k$ for all $k \in \mathcal{I}\backslash\{i\}$, there must be exactly one victorious coalition $\mathcal{C}$ and $\mathcal{I}_R$ must be a subset of $\mathcal{C}$. For the remainder of the proof, let p be the price vector announced by $\mathcal{C}$. If $\#\mathcal{I}_R \geq 2$, then $\mathcal{C}$ must be unanimous since otherwise the agents in $\mathcal{C}$ play an integer game with no equilibrium – each $i \in \mathcal{I}_R$ would have to announce a n^i such that $n^i > n^k$ for all $k \in \mathcal{C}\backslash\{i\}$. Continuing with the case where $\mathcal{C}$ is victorious (and hence unanimous) and $\#\mathcal{I}_R \geq 2$, the single deviation rule implies, for $i \in \mathcal{I}_R$, that x^i must be a $\succsim^i$-maximum on $\{w^i \in \mathbb{R}^L_+ : p \cdot w^i = p \cdot e^i,\ w^i \leq \sum_{k\in\mathcal{C}} e^k\}$. Given the monotonicity of $\succsim^i$, $p \gg 0$. Given interiority, the outcome for $\mathcal{I}_R \subset \mathcal{C}$, $(x^i)_{i\in\mathcal{I}_R}$, is strictly greater than 0 in every coordinate; therefore, since $\#\mathcal{I}_R \geq 2$, $\sum_{k\in\mathcal{C}} e^k \gg x^i$ for $i \in \mathcal{I}_R$. There must therefore be a $\varepsilon > 0$ such that any w^i with $\|x^i - w^i\| < \varepsilon$ and $p \cdot w^i = p \cdot e^i$ satisfies $w^i \leq \sum_{k\in\mathcal{C}} e^k$ and hence $x^i \succsim^i w^i$. The convexity of $\succsim^i$ then implies that if $\widetilde{x}^i \succsim^i x^i$ then $p \cdot \widetilde{x}^i \geq p \cdot e^i$. For if there were a $\widetilde{x}^i \in \mathbb{R}^L_+$ with $\widetilde{x}^i \succsim^i x^i$ and $p \cdot \widetilde{x}^i < p \cdot e^i$ then by convexity $\overline{x}^i = \lambda\widetilde{x}^i + (1-\lambda)x^i \succsim^i x^i$ for any $\lambda \in (0,1)$; and so by choosing λ sufficiently small and since $p \gg 0$, we can find a $w^i \gg \overline{x}^i$ with $\|x^i - w^i\| < \varepsilon$ and $p \cdot w^i = p \cdot e^i$, which by monotonicity satisfies $w^i \succ^i \overline{x}^i$ and hence $w^i \succ^i x^i$. So $\widetilde{x}^i \succsim^i x^i \Rightarrow p \cdot \widetilde{x}^i \geq p \cdot e^i$. But $\widetilde{x}^i \succ^i x^i$ and $p \cdot \widetilde{x}^i = p \cdot e^i$ cannot occur: if it did then $\widetilde{x}^i \succ^i x^i \succsim^i e^i$ and interiority give $\widetilde{x}^i \gg 0$ and so, by continuity, for any $\alpha \in (0,1)$ sufficiently near 1, $\alpha\widetilde{x}^i \succ^i x^i$ and $p \cdot \alpha\widetilde{x}^i < p \cdot e^i$. Hence, for $i \in \mathcal{I}_R$, x^i is $\succsim^i$-maximizing on $\{w^i \in \mathbb{R}^L_+ : p \cdot w^i = p \cdot e^i\}$. We can then apply the proof of Proposition 1, using the price p announced by all $k \in \mathcal{C}$, to conclude that the equilibrium satisfies Pareto efficiency for the rational agents. In the case where $\mathcal{C}$ is victorious and $\#\mathcal{I}_R = 1$, the agent $i \in \mathcal{I}_R$ must receive an outcome $x^i \succsim^i e^i$ since i could receive e^i by announcing $(\{i\}, n^i, p^i, e^i; a, 0, 1)$, where $n^i > n^k$ for all $k \in \mathcal{I}\backslash\{i\}$ (as at the beginning of the paragraph). Hence the equilibrium again satisfies Pareto efficiency for the rational agents. Finally, notice that if $\mathcal{I}_{IR} = \varnothing$ then $\mathcal{C} = \mathcal{I}$. Since, furthermore, the outcome x^i is $\succsim^i$-maximizing on $\{w^i \in \mathbb{R}^L_+ : p \cdot w^i = p \cdot e^i\}$ for each $i \in \mathcal{I}$, (p, x) must be an exchange equilibrium. So, when $\mathcal{I}_{IR} = \varnothing$, the outcome x of any equilibrium is the allocation of an exchange equilibrium. ■

Proof of Proposition 5. In the text. ■

References

[1] Alchian, A., 1950, 'Uncertainty, evolution and economic theory,' *Journal of Political Economy* 58: 211–221.

[2] Becker, G., 1957, *The Economics of Discrimination*, University of Chicago Press: Chicago.

[3] Blume, L., and Easley, D., 1990, 'Implementation of Walrasian expectations equilibria,' *Journal of Economic Theory* 51: 207-227.

[4] Blume, L, and Easley, D., 2006, 'If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets,' *Econometrica* 74: 929–966.

[5] Camerer, C., and Fehr, E., 2006, 'When does "economic man" dominate social behavior?' *Science* 311: 47-52.

[6] Dasgupta, P., Hammond, P., and Maskin, E., 1979, 'The implementation of social choice rules: some general results on incentive compatibility,' *Review of Economic Studies* 46: 185-216.

[7] De Long, J., Shleifer, A., Summers, L., and Waldmann, R., 1990, 'Noise trader risk in financial markets,' *Journal of Political Economy* 98: 703-38.

[8] Eliaz, F., 2002, 'Fault tolerant implementation,' *Review of Economic Studies* 69: 589-610.

[9] Feddersen, T., and Pesendorfer, W., 1996, 'The swing voter's curse,' *American Economic Review* 86: 408-424.

[10] Fehr, E., and Tyran, J.-R., 2005, 'Individual irrationality and aggregate outcomes,' *Journal of Economic Perspectives* 19: 43-66.

[11] Friedman, M., 1953, 'The methodology of positive economics' in M. Friedman (ed.), *Essays in Positive Economics*, University of Chicago Press: Chicago, 3–43.

[12] Haltiwanger, J., and Waldman, M., 1985, 'Rational expectations and the limits of rationality: an analysis of heterogeneity,' *American Economic Review* 75: 326-40.

[13] Haltiwanger, J., and Waldman, M., 1989, 'Limited rationality and strategic complements: the implications for macroeconomics,' *Quarterly Journal of Economics* 104: 463-83.

[14] Harrison, J. and Kreps, D., 1979, 'Martingales and arbitrage in multiperiod securities markets.' *Journal of Economic Theory* 20: 381-408.

[15] Hurwicz, L., 1972, 'On informationally decentralized systems,' in C. McGuire and R. Radner (eds.), *Decision and Organization*, North Holland: Amsterdam, 297-336.

[16] Hurwicz, L., 1979, 'On allocations attainable through Nash equilibria,' *Journal of Economic Theory* 21: 140–165.

[17] Hurwicz, L., Maskin, E., and Postlewaite, A., 1995, 'Feasible Nash implementation of social choice rules when the designer does not know endowments or production sets,' in J. Ledyard (ed.), *The Economics of Informational Decentralization: Complexity, Efficiency and Stability (essays in honor of Stanley Reiter)*, Kluwer Academic Publishers: Boston, p. 367-433.

[18] Jackson, M., Palfrey, T., and Srivastava, S., 1994, 'Undominated Nash implementation in bounded mechanisms,' *Games and Economic Behavior* 6: 474-501.

[19] Klette, T., and Kortum, S., 2004, 'Innovating firms and aggregate innovation,' *Journal of Political Economy* 112: 986–1018.

[20] Lentz, R., and Mortensen, D., 2005, 'Productivity growth and worker reallocation,' *International Economic Review* 46: 731-749.

[21] McFadden, D., 1969, 'A simple remark on the second best Pareto optimality of market equilibria,' *Journal of Economic Theory* 1: 26-38.

[22] McKelvey, R., and Palfrey, T., 1995, 'Quantal response equilibria for normal form games,' *Games and Economic Behavior* 10: 6-38.

[23] Nelson, R., and Winter, S., 1982, *An Evolutionary Theory of Economic Change*, Belknap Press: Cambridge.

[24] Postlewaite, A., and Schmeidler, D., 1986, 'Implementation in differential information economies,' *Journal of Economic Theory* 39: 14-33.

[25] Samuelson, P., 1965, 'Proof that properly anticipated prices fluctuate randomly,' *Industrial Management Review* 6: 41-49.

[26] Sandroni, A., 2000, 'Do Markets Favor Agents Able to Make Accurate Predictions?' *Econometrica* 68: 1303-1342.

[27] Satterthwaite, M., and Sonnenschein, H., 1981, 'Strategy-proof allocation mechanisms at differentiable points,' *Review of Economic Studies* 48: 587-597.

[28] Serizawa, S., 2002, 'Inefficiency of strategy-proof rules for pure exchange economies,' *Journal of Economic Theory* 106: 219-241.

[29] Sutton, J., 1997, 'One smart agent,' *RAND Journal of Economics* 28: 605-628.

[30] Zhou, L., 1991, 'Inefficiency of strategy-proof allocation mechanisms in pure exchange economies,' *Social Choice and Welfare* 8: 247-254.