Dogra, Keshav

Working Paper

Paradoxes and problems in the causal interpretation of equilibrium economics

Staff Report, No. 1093

Provided in Cooperation with:
Federal Reserve Bank of New York

Suggested Citation: Dogra, Keshav (2024) : Paradoxes and problems in the causal interpretation of equilibrium economics, Staff Report, No. 1093, Federal Reserve Bank of New York, New York, NY, https://doi.org/10.59576/sr.1093

This Version is available at:
https://hdl.handle.net/10419/300472

STAFF
REPORTS

NO. 1093
MARCH 2024

Paradoxes and Problems in the Causal Interpretation of Equilibrium Economics

Keshav Dogra

FEDERAL RESERVE BANK *of* NEW YORK

Paradoxes and Problems in the Causal Interpretation of Equilibrium Economics

Keshav Dogra

Federal Reserve Bank of New York Staff Reports, no. 1093
March 2024
https://doi.org/10.59576/sr.1093

Abstract

Equilibrium assumptions posit relations between different people's beliefs and behavior without describing a process that causes these relations to hold. I show that because equilibrium models do not describe a causal process whereby one endogenous variable affects another, attempts to decompose the effects of shocks into "direct" and "indirect" effects can suggest misleading predictions about how these models work. Equilibrium assumptions also imply absurd paradoxes: history can determine future behavior without affecting any intervening state variables today; individuals can learn information that no one originally possesses by observing each other's actions. This makes equilibrium models unreliable tools to study how economic systems coordinate activity and aggregate dispersed information. I describe how to construct non-equilibrium models that avoid these paradoxes and can be interpreted causally.

JEL classification: B41, C70, D50, D83, E70
Key words: equilibrium, disequilibrium, mechanisms, causality, paradoxes

---

Dogra: Federal Reserve Bank of New York (email: keshav.dogra@ny.frb.org). The author thanks Sushant Acharya, James Best, Stéphane Dupraz, Colin Hottman, and Patrick Sun for helpful comments and for discussions over many years which have informed the ideas presented in this paper.

1 Introduction

Economic theory is built on **equilibrium** assumptions. Perhaps this term once connoted the rest point of a dynamic process. But in modern parlance, an equilibrium assumption posits a relation between the beliefs and behavior of different agents without explicitly describing a process which causes this relation to hold. Competitive equilibrium assumes prices clear markets, without specifying who sets prices, or how they find a market-clearing level; Nash equilibrium assumes each player's strategy is optimal given others', without describing how these strategies came to be consistent; rational expectations assumes agents' subjective probability distributions over all variables, including others' behavior, equal objective distributions. There have been various attempts to support these assumptions by showing that an explicit disequilibrium process converges to equilibrium, with fairly mixed results. But such concerns are marginal today; most economists are happy to assume equilibrium directly, and treat the question of stability, if at all, in a handwaving fashion.

At the same time, we intuitively think about the world in terms of causal processes, whereby a cause triggers a chain of other events culminating in its effect. Formalizing this intuition – identifying the mechanisms through which causes produce their effects – is central to science, including applied economics. It is not surprising, then, that we try to apply the same 'causal process' framework to understand and interpret equilibrium models. We search for intuitive, 'process' stories to explain our models' results; evaluate their behavioral assumptions (e.g. rational expectations) based on whether these plausibly describe people's decision processes; and even try to 'inspect the mechanism' by decomposing the effects of shocks and policies into direct and indirect effects. Taken literally, there is an inconsistency between the 'causal process' view, in which one endogenous variable affects another through a causal chain, and equilibrium models, in which all endogenous variables are determined simultaneously. But perhaps this is not a concern: equilibrium models might be consistent with *some* causal process, even if they don't spell it out; and besides, the stories we tell about models are just aids for intuition, not worth taking too seriously.

I will argue that, on the contrary, this inconsistency poses serious problems. Section 2 proposes intuitive principles which any model providing a complete description of causal process should satisfy. In models satisfying these principles, understanding mechanisms is useful: it helps generalize results, design interventions, and assess the sensitivity of results to parameters. But equilibrium models violate these principles, and cannot be consistently interpreted in terms of causal process: there are no 'mechanisms' to understand. Thus, as I show in Section 3, tools such as direct-indirect effects decompositions (Kaplan et al., 2018) do not help understand such models, and can provide misleading conclusions about generalization, designing interventions, and parameter sensitivity.

Equilibrium models are not just hard to interpret causally; they also produce absurd predictions which could not arise from any plausible process. To illustrate this, I describe two paradoxes. In the first paradox, which arises in extensive form games, assuming equilibrium rules out outcomes which intuitively should be possible. In my leading example, agents freely choose prices for their output in the morning, and consume and produce in the afternoon. The assumption that agents set prices optimally in the morning selects a unique level of output ('full employment') in the afternoon. But

if prices are *fixed* at the same level that they take in flexible-price equilibrium, *any* level of output is an an equilibrium of the afternoon subgame. How can price flexibility in the morning ensure full employment in the afternoon, if not by affecting what prices are actually set? Section 4 explains how such paradoxes arise, and argues that they should lead us to distrust equilibrium assumptions. This is not a call for refinements to select among multiple equilibria, ruling out the 'unreasonable' ones. Equilibrium assumptions *themselves* rule out reasonable outcomes, and should be relaxed.

Section 5 discusses the second paradox, which arises in rational expectations equilibrium (REE): when agents learn from endogenous variables (e.g. prices), these variables can reveal information that no one is endowed with (Dubey et al., 1987). Suppose a number of agents each seek to set their action equal to the same fundamental θ. Each observes others' average action, but none observe θ. There is a valid REE in which everyone's action equals θ; each agent i simply sets her action equal to the average action, knowing that, in equilibrium, this perfectly tracks θ. This is absurd: the information can never 'get into' actions if no one knows it to begin with. Again, the problem is not easily removed with refinements: sometimes *all* REEs feature 'immaculate revelation'.

The key to both paradoxes is that the equilibrium assumption that an agent acts optimally does not, as one might think, describe a process through which her action is causally determined by her beliefs and the environment prior to her decision. Instead, it describes a non-causal, 'simultaneous' relation between various agents' current and future beliefs and actions, which does not respect the arrow of time. There is no guarantee that any plausible process leads this relation to hold.

Several alternatives to full information rational expectations have been proposed: rational inattention, diagnostic expectations, cognitive discounting, etc. These have been motivated, in part, by the desire to provide a more realistic description of the way people form beliefs. But the problem with REE is not that it posits an *unrealistic* belief formation process; it is that it assumes a relation between different agents' beliefs and behavior without describing *any* process causing this to hold. The 'alternatives' make no progress in this respect, since they are still equilibrium models: they now assume a 'distorted' relation between beliefs and behavior, but still do not describe a process causing this to hold. Thus, as I show in Section 6, they exhibit the same paradoxes as REE.

A better alternative is to construct *process models*, which explicitly describe how all endogenous variables are determined by agents through a recursive causal process. This approach has many precursors. Perhaps the closest is the 'sequence analysis' of the Stockholm School, in which agents enter date t with potentially inconsistent ex ante expectations and plans; their attempt to execute these plans leads to ex post outcomes which disappoint some agents' expectations, leading to revision of date $t+1$ plans and expectations; etc.[1] Market games (e.g. Shapley and Shubik (1977); see Giraud (2003) for a review) approach process models, by describing explicitly how endogenous variables like prices are determined by individual action, but still assume Nash equilibrium. Conversely, the temporary equilibrium (Grandmont, 1977) and adaptive learning (Evans and Honkapohja, 2001) literatures generally assume spot market clearing, but depart from rational expectations by assuming a learning rule. While both remain equilibrium approaches, combining them – studying learning in an explicit market game, as I do in Section 4.4 – is one way to construct

[1] See Lundberg (1937) and Lindahl (1939) for two central contributions and Hansson (1982) for a survey.

process models.[2] Such models avoid the above paradoxes, and their mechanisms can be understood.

This paper's thesis – assuming equilibrium is groundless unless we can explicitly describe a plausible process causing this assumption to hold – echoes several older literatures.[3] Relative to these literatures, the paper makes three contributions. First, I identify a new implausible implication of equilibrium assumptions in a class of dynamic games. This casts doubt on equilibrium assumptions generally, and more specifically on the "classical" thesis, inherited by New Keynesian models, that flexible prices ensure full employment. While both this and 'immaculate revelation' might seem like technical curiosities, whether and how markets avoid aggregate demand failures (Keynes, 1936), or aggregate dispersed information (Hayek, 1945), are core questions in economics. If equilibrium models cannot be trusted to answer them, we need alternatives. Second, while direct-indirect effect decompositions have become popular in the macroeconomic literature, there is surprisingly little discussion of what they are useful for. I suggest a way to evaluate their usefulness, and find that they can be highly misleading in equilibrium models.[4] Finally, my formal distinction between equilibrium and process models connects various literatures, highlights the common source of various problems with equilibrium models, and identifies which of the many possible deviations from benchmark models such as REE will avoid these problems. Next, I describe this distinction.

2 Defining process and equilibrium models

While there are important differences between competitive equilibrium, Nash equilibrium, rational expectations, etc., this paper argues that all such concepts provide incomplete descriptions of causal process, leading to similar problems. To that end, I now provide general definitions of both 'causally complete' and equilibrium models, formalizing the notion that equilibrium conditions posit relations between different agents' behavior without describing a process causing these relations to hold.[5]

A **model** is a set of agents A, endogenous variables $\boldsymbol{y} = \{y_i\}_{i \in I} \in Y$ with index set I, exogenous variables $\boldsymbol{z} = \{z_j\}_{j \in J} \in Z$ with index set J, and 'conditions' indexed by $k \in K$ and described by correspondences $\Gamma_k : X_{R_k} \rightrightarrows Y_{L_k}$, where $X = Y \times Z$, $R_k \subseteq I \cup J$, $L_k \subseteq I$. That is, the model's kth condition states that $\boldsymbol{y_{L_k}} \in \Gamma_k(\boldsymbol{x_{R_k}})$. The model's **reduced form** is the set of possible outcomes $\boldsymbol{y}$ satisfying all these conditions given $\boldsymbol{z}$, $Y^*(\boldsymbol{z}) := \{\boldsymbol{y} \in Y | \boldsymbol{y_{L_k}} \in \Gamma_k(\boldsymbol{x_{R_k}}), \forall k \in K\}$.

Example: competitive equilibrium of a pure exchange economy. There are ℓ goods

[2] See Godley and Lavoie (2016) for another way. My definition of process models also relates to Tesfatsion (2017)'s definition of agent-based computational economics (ACE). Process models need not *necessarily* be solved using the ACE approach, but doing so is always *feasible*, without the fixed-point calculations required in equilibrium models.

[3] Adequately describing these important literatures is well beyond the scope of this paper. See Fisher (1983) for a seminal contribution and Schinkel (2001) for a survey. On learning in games, see Hart and Mas-Colell (2003); Young (2004); on learning in macroeconomics, Lindh (1989); Eusepi and Preston (2023); for evolutionary approaches, Howitt and Clower (2000); Gintis (2007); Mandel and Gintis (2016).

[4] The mediation analysis literature (e.g. Judd and Kenny (1981); VanderWeele (2015)) does discuss how decompositions are useful, but mostly considers recursive models. The smaller literature studying nonrecursive models (e.g. Bollen (1987)), does not, to my knowledge, ask whether decompositions are equally useful in such settings.

[5] In other disciplines, and in economics prior to around 1945, 'equilibrium' denotes a rest point of a dynamic process. That is not the sense in which the word is used in modern economics or in this paper. Sometimes economists use 'equilibrium' to mean any 'solution' of a model, i.e. any outcome logically consistent with its assumptions. I will not use that terminology. The premises of a process model entail an outcome, but that outcome is not an equilibrium.

and m consumers with utility functions $u_i : \mathbb{R}^\ell_{++} \to \mathbb{R}$ and endowments $\boldsymbol{\omega_i} \in \mathbb{R}^\ell_+$. The agents are the consumers, endogenous variables are $\{\boldsymbol{x_i}\}_{i=1}^m$ and $\boldsymbol{p} \in \mathbb{R}^\ell_+$, and exogenous variables are $\boldsymbol{\omega} = \{\boldsymbol{\omega_i}\}_{i=1}^m$. The model has two sets of conditions. First, for each $i = 1, ..., m$, $\boldsymbol{x}_i$ maximizes $u_i(\boldsymbol{x})$ in $\mathbb{R}^\ell_{++}$ subject to the budget constraint $\boldsymbol{p} \cdot (\boldsymbol{x} - \boldsymbol{\omega}_i) \leq 0$. We can write this as the correspondence

$$\boldsymbol{x_i} \in \Gamma_i(\boldsymbol{p}, \boldsymbol{\omega_i}) := \arg \max_{\boldsymbol{x} \in \mathbb{R}^\ell_{++}} u(\boldsymbol{x}) \text{ s.t. } \boldsymbol{p} \cdot (\boldsymbol{x} - \boldsymbol{\omega_i}) \leq 0 \tag{1}$$

Second, the market for each good $j = 1, ..., \ell$ clears, $\sum_{i=1}^m (x_i^j - \omega_i^j) = 0$. We can write these remaining conditions $m + 1, ..., m + \ell$ as the correspondences

$$\{x_i^j\}_{i=1}^m \in \Gamma_{m+j}(\omega) := \left\{ \{x_i^j\}_{i=1}^m \middle| \sum_{i=1}^m (x_i^j - \omega_i^j) = 0 \right\} \tag{2}$$

The reduced form is the set of equilibria $(\{\boldsymbol{x_i}\}_{i=1}^m, \boldsymbol{p})$ satisfying all $m + \ell$ conditions, given $\boldsymbol{\omega}$.

Some conditions in economic models are **behavioral**. In the previous example, intuitively, the first m conditions (1) describe how agents 1 through m behave, while the remaining ℓ conditions (2) do not describe how any agents behave. Whether a condition k is behavioral is not a mathematical property of the model, but depends on its economic interpretation. A necessary condition is that its left-hand-side variables L_k are all chosen by the agent, but this is not sufficient. In a single-agent version of the last example ($m = 1$), each market clearing condition has the agent's consumption of some commodity j as its left-hand-side variable, but these conditions do not describe how they behave. We simply take the fact that some conditions are identified with particular agents' behavior as a primitive of the model. In most cases, these conditions will describe optimizing behavior.

A **process model** is a model satisfying the following principles:

1. The directed graph on I constructed by drawing an edge from i' to i if $i' \in R_k, i \in L_k$ for some k is acyclic and respects the arrow of time: variables in R_k temporally precede L_k.

2. Each correspondence Γ_k is non-empty-valued.

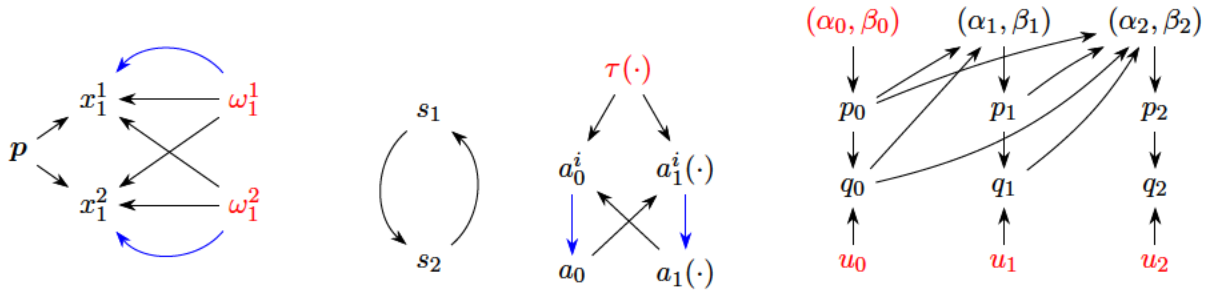
3. Each endogenous variable $i \in I$ appears as the output of exactly one condition, $i \in L_k$.

4. If i is not determined by legal/institutional rules, accounting identities, or non-human physical processes, it is chosen by an agent a according to a behavioral condition k. L_k only includes a's choice variables. R_k only includes variables known or experienced by a before choosing i.

A process model consists of conditions recursively (1) describing how each endogenous variable is determined by rules, identities, physical processes or individual action (4). One condition determines each variable (3) and produces a well-defined outcome for any input (2). These are intuitively reasonable principles that any model attempting to provide a complete description of causal process should satisfy. This paper shows that violating these principles can lead to problems.

We define an **equilibrium model** negatively, as a model containing behavioral conditions which violates one or more of principles 1-4. Our exchange economy violates 3 and 4; Figure 1a

Figure 1: Graphs of three equilibrium models and one process model



(a) Competitive equilibrium (b) Nash (c) REE (d) Learning about demand

depicts the associated graph when $m = 1$, $j = 2$.[6] Each agent's demand x_i^j is the output of both condition i (i's optimal consumption, shown by black edges in Figure 1a) and condition $m + j$ (market clearing for good j, shown by the blue edges), violating 3. Prices $\boldsymbol{p}$ are endogenous, yet are not determined by rules, identities, physical processes or behavioral conditions, violating 4.

Example: Nash equilibrium. There are 2 players with payoffs $u_i(s_i, s_{-i})$ and strategy sets S_i. The agents are $1, 2$ and the endogenous variables are (s_1, s_2); there are no exogenous variables. There are 2 conditions, both behavioral: for $i = 1, 2$, $s_i \in \arg\max_{s \in S_i} u_i(s, s_{-i})$. This is not a process model since the correspondences form a cyclic graph, violating Principle 1 (see Figure 1b). If player 1 chooses s_1 before 2 chooses s_2, the graph also violates the arrow of time, since s_1 depends on but also temporally precedes s_2. Finally, if $S_i = \mathbb{R}$, $u_i(s_i, s_{-i}) = s_i s_{-i}$, the model also violates Principle 2, since $\Gamma_i(s_{-i}) = \arg\max_{s \in \mathbb{R}} s s_{-i}$ is empty-valued for $s_{-i} \neq 0$.

Exogenous variables $\boldsymbol{z}$ include both shocks and policy interventions. In dynamic rational expectations models, an intervention cannot be specified as an 'arbitrary' sequence of forcing variables $\{z_t\}$, since optimal behavior depends on the objective distribution of z_t (Lucas, 1976). This can be accommodated within our definition by specifying $\boldsymbol{z}$ as a *function* of the stochastic fundamentals.

Example: Rational expectations. A continuum of agents $i \in [0, 1]$ act at dates $t = 0, 1$. At date 1, Nature draws $\theta \sim N(0, 1)$ and agents choose $a_1^i = \theta + \tau(\theta) + \delta a_0$. At date 0, the government commits to a policy $\tau(\theta)$, and agents choose $a_0^i = \alpha \mathbb{E}_0 \tau(\theta) + \beta \mathbb{E}_0 a_1$. Aggregate actions are $a_t = \int_0^1 a_t^i di$, $t = 0, 1$. The endogenous variables are scalars a_0, $(a_0^i)_{i \in [0,1]}$ and functions $a_1(\cdot)$, $(a_1^i(\cdot))_{i \in [0,1]}$, the exogenous variable is the function $\tau(\cdot)$, and the model conditions, shown in Figure 1c, are

$$a_0^i = \alpha \int \tau(\theta) d\Phi(\theta) + \beta \int a_1(\theta) d\Phi(\theta),\ a_1^i(\theta) = \theta + \tau(\theta) + \delta a_0,\ a_0 = \int_0^1 a_0^i di,\ a_1(\theta) = \int_0^1 a_1^i(\theta) di$$

This is not a process model: it has a cycle and an edge $a_1(\theta) \to a_0^i$ which violates time's arrow.

Whether a model is equilibrium or process depends on its 'structural' representation (i.e. all the model's conditions), not on its reduced form. The reduced form of an equilibrium model may not be distinguishable from that of a process model. Indeed, given an equilibrium model, one may be

[6] In these figures, exogenous variables are highlighted in red, endogenous variables are shown in black, behavioral conditions are depicted by black edges, and other conditions are depicted by blue edges.

able to construct a process model which has the same reduced form. In the last example, directly assuming agents choose $a_0^i = (1-\beta\delta)^{-1}(\alpha+\beta)\mathbb{E}\tau(\theta)$, instead of assuming they act based on rational expectations about date 1 outcomes, yields a valid process model with the same reduced form.

The assumption of optimizing behavior is not necessarily an equilibrium assumption: it depends what variables agents take as given when optimizing. Assuming agents optimize given variables determined simultaneously, or in the future, violates Principle 1. But assuming agents optimize given beliefs which are determined prior to their decision satisfies this principle. Thus, a well-specified game can be converted into a process model by replacing equilibrium conditions with the weaker assumption of optimization given beliefs, together with a rule describing belief formation.

Example: learning about demand. A monopolistic firm sets prices to maximize expected profit given the perceived demand curve $\alpha_t - \beta_t p_t$, and produces as necessary to meet demand at cost $q^2/2$. At time t the firm sets $p_t = \max_{p\geq 0} p[\alpha_t - \beta_t p] - \frac{1}{2}[\alpha_t - \beta_t p]^2$. Actual demand is $q_t = a - bp_t + u_t$ where u_t is white noise. The firm updates its estimate of the demand curve using least squares: $(\alpha_{t+1}, \beta_{t+1}) = \arg\min_{\alpha\geq 0, \beta\geq \underline{\beta}} \sum_{\tau=0}^{t}(q_\tau - \alpha + \beta p_\tau)^2$ where $\underline{\beta} > 0$. This is a process model. The agents are the firm (who chooses p_t, α_t, β_t) and consumers (who choose q_t); endogenous variables, $\{\alpha_{t+1}, \beta_{t+1}, p_t, q_t\}_{t=0}^{\infty}$; exogenous variables, $\alpha_0, \beta_0, \{u_t\}_{t=0}^{\infty}$. The model's conditions map $\{q_\tau, p_\tau\}_{\tau=0}^{t-1}$ to (α_t, β_t), (α_t, β_t) to p_t, and (p_t, u_t) to q_t. The associated graph (Figure 1d) respects the arrow of time, with each endogenous variable on the left hand side of exactly one condition.

My distinction between process and equilibrium models combines two distinct criteria. Process models must have a recursive structure (Principles 1-3); and they must describe how endogenous variables like prices are determined by human action (Principle 4). Principles 1-3 draw on Bentzel and Hansen (1954); Wold (1954); Strotz and Wold (1960)'s distinction between recursive and non-recursive systems, with a few differences.[7] First, my definition of process models permits correspondences, rather than functions, to allow for the possibility that we can only make set predictions about some outcomes (e.g. how an individual will choose between alternatives when indifferent). We can still solve for the set of possible outcomes $Y^*(\mathbf{z})$ recursively. Take all conditions K_0 with only exogenous variables as inputs, $R_k \subseteq J$; select any $\boldsymbol{y}_{\boldsymbol{L_k}} \in \Gamma_k(\boldsymbol{x}_{\boldsymbol{R_k}})$ for each $k \in K_0$. Then take all conditions $K_1 = K \setminus K_0$ with only exogenous and previously-determined endogenous variables as input; select any $\boldsymbol{y}_{\boldsymbol{L_k}} \in \Gamma_k(\boldsymbol{x}_{\boldsymbol{R_k}})$ for each $k \in K_1$. If at some stage in the process multiple values of $\boldsymbol{y}_{\boldsymbol{L_k}}$ are possible, these all remain possible: if L_k's direct causes R_k do not uniquely determine these variables, nothing else can. As Section 4 describes, this is not true in equilibrium models.

The second difference is that process models are block-recursive (in Strotz and Wold (1960)'s terminology, 'vector causal'), not strictly recursive: $\boldsymbol{x}_{\boldsymbol{R_k}}$ causes $\boldsymbol{y}_{\boldsymbol{L_k}}$, but no causal relations among the variables within $\boldsymbol{y}_{\boldsymbol{L_k}}$ are defined. This is a natural way to model multivariate decision problems: in the example above, (α_t, β_t) are jointly determined by $\{q_\tau, p_\tau\}_{\tau=0}^{t-1}$, but it would not make sense to ask how α_t causally affects β_t. But every model, considered as a single block, is 'vector causal' in the

[7] See also Bentzel and Wold (1946). While Wold was also engaged in a debate about how to *estimate* models – which is beyond the scope of my paper – his distinction was largely intended to clarify the difference between the *causal interpretation* of Stockholm School 'process analysis' models (including Tinbergen's macroeconometric models), and Haavelmo and the Cowles Commission's 'interdependent systems' (Morgan, 1991; Richardson, 1996).

trivial sense that all exogenous variables $\boldsymbol{z}$ cause $\boldsymbol{y}$. Thus, requiring a model's conditions to be block-recursive is vacuous without stipulating how a model's assumptions can be partitioned into distinct 'conditions'. I do so by requiring that each condition k which does not describe rules, identities or physical processes describes the behavior of a single model agent (Principle 4). Intuitively, equilibrium models implicitly or explicitly assume relations between different agents' behavior; process models make separate assumptions about each agent, and study their implications.

To see why we need Principle 4, consider two models. First, a firm chooses inputs $x_1, ..., x_\ell$ to maximize profit, given a smooth decreasing returns production function f and input prices $\boldsymbol{w} \in \mathbb{R}^\ell_+$. Denoting the ith factor's marginal product by f_i, the firm's optimal factor demand is

$$\boldsymbol{x} \in \Gamma(w) := \{\boldsymbol{x} \in \mathbb{R}^\ell_+ | f_i(x_1, ..., x_\ell) = w_i, i = 1, ..., \ell\} \tag{3}$$

This is a process model: the endogenous variables $x_1, ..., x_\ell$ are determined as the output of a non-empty-valued behavioral correspondence. Now consider an exchange economy, where prices $p_1, ..., p_\ell$ are determined by ℓ market clearing conditions $d_i(p_1, ..., p_\ell) = \omega^\ell$ where d_i denotes demand for commodity i, generated by a consumer with quasilinear preferences. We can represent this as

$$\boldsymbol{p} \in \Gamma(\omega) := \{\boldsymbol{p} \in \mathbb{R}^\ell_+ | d_i(p_1, ..., p_\ell) = \omega_i, i = 1, ..., \ell\} \tag{4}$$

Intuitively, (4) is an equilibrium model, since it does not describe a process by which prices come to equate demand and supply; while (3) is a process model, since it does describe how firms choose input demands to maximize profit. But mathematically, (3) and (4) are identical: in one case all input demands are block-recursively determined by all factor prices, in the other all prices are determined by all endowments. The difference is not mathematical, but substantive: input demands are determined by the firm in the first model, while prices are not determined by any agent in the second model ((4) does not describe the behavior of any individual).

To be clear, I am not arguing there is any merit in introducing an 'auctioneer' who chooses prices to minimize excess demand, making (4) a behavioral condition and producing a process model (albeit an unrealistic one). Rather, my point is that when making assumptions on endogenous variables, it is useful to spell out exactly what those entail about individual behavior. It is easier to interrogate implausible assumptions like the 'auctioneer' if they are made explicitly.

Next, I explain why process models, which satisfy Principles 1-4, are easier to understand and interpret causally than equilibrium models, which violate these principles.

3 'Understanding mechanisms' in equilibrium models

The study of mechanisms – the process by which a cause produces its effect – is central to diverse scientific disciplines.[8] So we might expect the study of mechanisms to play a similarly central role

[8]To name a few: cell biology (Bechtel, 2006), neuroscience (Craver, 2007), the study of disease (Thagard, 2000), program evaluation (Pawson and Tilley, 1997), and the social sciences (Hedström and Ylikoski, 2010; MacKinnon, 2012). The voluminous philosophical literature on mechanisms includes Glennan (1996); Machamer et al. (2000).

in economics.[9] And indeed, economists often use causal language when describing the effect of one endogenous variable on another. Discussion of the 'monetary transmission mechanism' (Mishkin, 1995) decomposes the effect of monetary policy on the economy into a 'real interest rate channel', 'exchange rate channel', etc.; this seems to describe a process in which monetary policy first affects exchange rates, credit supply, etc., and then each of these factors affects economic activity. Formal decompositions into such channels have become popular in the heterogeneous agent New Keynesian (HANK) literature (Kaplan et al., 2018). But this 'causal process' language does not correspond to the actual structure of equilibrium models, which do not describe a causal chain whereby one endogenous variable affects another; rather, all equations simultaneously determine all endogenous variables. Thus, I will argue, attempts to decompose the mechanisms through which exogenous variables produce their effects generally do not help us understand equilibrium models.

This invites two questions. First, how can you show whether a tool helps understand a model – isn't this subjective? Second, why should we care? Once we've solved a model, why is 'understanding' the result useful; and if it is, can't other tools besides decompositions facilitate understanding?

While understanding has a pragmatic dimension, it is not purely subjective: I may *feel* I understand something while failing to do so (De Regt, 2017). But it is hard to define. Indeed, while the HANK literature often argues that decompositions help 'understand a model's mechanisms', it rarely spells out exactly what decompositions are useful for, or what counts as 'understanding'. I will adopt an operational definition. Suppose we know how the endogenous variables $\boldsymbol{y}$ depend on exogenous variables $\boldsymbol{z}$ in a model given a particular 'setting'. We often want to know how $\boldsymbol{y}$ changes under 'perturbations' to this setting. If our model delivers implausible results, which assumptions must be changed to fix them? If someone else's model has a surprising result, is it sensitive to assumptions? Even if we trust a model's conclusions, do the results generalize to other theoretical or (the most radical 'perturbation' of all) real-world settings? If $\boldsymbol{z}$ is a policy intervention, can we design alternative policies with the same benefits but fewer side-effects?

Models have many assumptions and parameters, so it would be costly to solve *every* possible model configuration; we must intuitively assess which perturbations seem promising, even if we check that intuition by solving a few alternative specifications. When assessing the robustness of someone else's conclusions, or generalizing from model to reality, solving the perturbed model is outright impossible. Hence my operational definition: if I understand the model, I should be able to qualitatively predict how $\boldsymbol{y}$ changes under perturbations to the model 'setting', without explicitly solving the 'perturbed' model.[10] Many tools can facilitate 'understanding' in this sense: supply-demand graphs, 'toy' models, etc. I will argue that in equilibrium models, a particular set of

[9] Deaton (2010) forcefully argues that studying mechanisms is essential to understanding economic development.

[10] This definition draws on both the mediation literature (see especially Judd and Kenny (1981, p603) for a clear statement of the practical benefits of 'process analysis') and elements of the philosophical literature which seem to me to capture what it means to 'understand' an economic model, in particular De Regt and Dieks (2005) ("a theory T is intelligible for scientists...if they can recognise qualitatively characteristic consequences of T without performing exact calculations") and Woodward (2003), who emphasizes the practical value of being able to answer *what-if-things-had-been-different* or *w-questions*. Philosophers broadly agree that understanding entails the ability to predict what would happen in various counterfactuals, but disagree as to whether understanding is constituted by, or the ground of, these abilities (see Hills (2016) and the various contributions in Grimm et al. (2016) for differing perspectives).

tools – attempts to decompose the total effect of $\boldsymbol{z}$ on $\boldsymbol{y}$ into its effect via various causal pathways – suggest misleading predictions about how $\boldsymbol{y}(\boldsymbol{z})$ would change under various perturbations. This involves some judgement – I cannot know for sure how another economist would predict – but I will show formally that decompositions are less informative in equilibrium than in process models.

3.1 Studying mechanisms in process and equilibrium models

A **process** model where every correspondence Γ_k is single-valued is a *structural causal model* (SCM) (Pearl, 2009), comprising n equations in n endogenous variables $y_1, ..., y_n$ and k exogenous variables $\mathbf{z} = z_1, ..., z_k$: $y_i = f_i(\mathbf{pa}_i)$, $i = 1, ..., n$, where $\mathbf{pa}_i \subseteq \{y_1, ..., y_{i-1}, \mathbf{z}\}$ (the 'causal parents' of y_i) are the variables that directly affect y_i. Since this system is recursive (cf. Principle 1), we can represent it by a directed acyclic graph (DAG), depicting variables by nodes, drawing a directed edge from each node of $\mathbf{pa}_i$ to its 'child node' y_i, where without loss of generality we 'topologically sort' endogenous variables (y_i can depend on y_{i-1}, but not vice versa). As in any structural model, we compute the effect of any exogenous z_j on any endogenous y_i by changing z_j and re-solving the system. But recursive models also allow us to define the effect of one *endogenous* variable y_j on another, y_i: we 'strike out' the jth equation, replace it with $y_j = \overline{y}_j$ (equivalently, treat y_j as exogenous), and solve the new system for a different value of $\overline{y}_j$ (Strotz and Wold, 1960). In Pearl (2009)'s terminology, we perform 'surgery' on the graph, removing all edges from $\mathbf{pa}_j$ to y_j.[11]

Figure 2: Direct and indirect effects in a structural causal model

(a) Direct effect of z on y (b) Indirect effect via pathway $z \to x_1 \to y$

One can also decompose the effect of exogenous variables on endogenous variables into indirect and direct effects.[12] Suppose a single exogenous variable z_1 (which we call z for simplicity) affects our outcome of interest y_n (henceforth y) both directly, and indirectly via mediator variables $y_1, ..., y_{n-1}$ (henceforth $\mathbf{x} = x_1, ..., x_{n-1}$). The direct effect of z on y is given by varying z while fixing $\mathbf{x}$, i.e. replacing the first $n-1$ equations by $x_i = \overline{x}_i, \forall i$, or removing all edges from the graph except the one from z to y (see Figure 2a). The indirect effect of z on y is defined residually as the total effect minus the direct effect; this can be further decomposed into the effect of z via various pathways (Albert and Nelson, 2011). Take any subset of the $n-1$ mediators, $i_1, ..., i_m$ where $1 \leq m \leq n-1$: this defines a path from z to x_{i_1} to x_{i_2}...to x_{i_m} to y. Removing all other edges, and computing the effect of z on y in the resulting subgraph, yields the indirect effect of z on y via the pathway $i_1 i_2 \cdots i_m$ (see Figure 2b for an example). There are $2^{n-1} - 1$ such pathway-specific

[11] While in principle one could also use potential outcomes (Rubin, 1974; Holland, 1986) notation, SCM terminology provides a more convenient way to *define* the effect of y_j on y_i in a recursive model. This is separate from the debate on the comparative advantage of the two approaches for causal inference in empirical economics (Imbens, 2020).

[12] This discussion follows closely the literature on the counterfactual approach to mediation analysis. See for example Robins and Greenland (1992), Pearl (2001), VanderWeele (2015), Albert and Nelson (2011).

indirect effects, which can be grouped in various ways, depending on the question of interest.

A generic **equilibrium** model cannot be represented as a recursive system mapping $\mathbf{z}$ to $\mathbf{y}$; instead, it must be represented as n *simultaneous* equations, $f_i(\mathbf{y}, \mathbf{z}) = 0$.[13] To define y_j's effect on y_i, we need a 'directional' equation expressing y_i as a function of y_j, where the equality sign stands for "is caused by" (Pearl (2009) p378).[14] But an equilibrium condition treats all endogenous variables symmetrically, without indicating which one is caused by the others and should appear on the left hand side. Worse, the same equilibrium conditions can generally be represented by different sets of equations. So it is ambiguous how one would use the methods above to compute the effect of one endogenous variable on another, or decompose the effect of z on y into different pathways.

Werning (2022) describes how naive manipulation of equilibrium conditions can lead to contradictory conclusions about the effect of expected inflation on realized inflation in New Keynesian models. Since inflation π_t and output gaps x_t satisfy the Phillips curve (NKPC) $\pi_t = \kappa x_t + \beta \mathbb{E}_t \pi_{t+1}$, one might think the effect of short-run inflation expectations $\mathbb{E}_t \pi_{t+1}$ on π_t is β (≈ 1 in standard calibrations), while long-run expectations play no role (Rudd, 2022). But as Werning points out, one could equally well solve the NKPC forward to get $\pi_t = \mathbb{E}_t \sum_{k=0}^{\infty} \beta^k x_{t+k}$ and conclude that inflation does not depend on expected *inflation* at any horizon, only on current and expected output gaps. Hazell et al. (2022) instead define $\widetilde{x}_{t+k} = x_{t+k} - \mathbb{E}_t x_{t+\infty}$ (where $\mathbb{E}_t x_{t+\infty}$ and $\mathbb{E}_t \pi_{t+\infty}$ denote long-run expectations of x_t and π_t) and write $\pi_t = \mathbb{E}_t \pi_{t+\infty} + \mathbb{E}_t \sum_{k=0}^{\infty} \beta^k \widetilde{x}_{t+k}$, arguing that *long-run* inflation expectations affect current inflation. We could even insist that the NKPC really describes how inflation determines *output gaps*, since $x_t = \kappa^{-1}(\pi_t - \beta \mathbb{E}_t \pi_{t+1})$. The NKPC is an equilibrium condition describing a relation which output gaps and actual and expected inflation must satisfy; it does not describe how any one of these endogenous variables is determined by the others.

To formally discuss mechanisms in equilibrium models, we need an unambiguous way of representing at least *some* equilibrium conditions as directional equations. Obvious candidates are the behavioral equations of individuals, which are naturally interpreted as describing how individual decisions (e.g. household consumption) are determined by variables the individual takes as given (e.g. prices and income).[15] One can decompose the effect of any exogenous variable on the individual's behavior into its effect on each of the variables they take as given. To illustrate this method and its limitations, I discuss the most influential recent example, Kaplan et al. (2018)'s analysis of the monetary policy transmission mechanism in representative agent (RANK) and HANK models.[16]

Consider the following RANK model. Time is discrete and there is perfect foresight. The representative household takes as given a sequence of real interest rates, wages and real transfers from monopolistically competitive firms $\{r_t, w_t, T_t\}_{t=0}^{\infty}$, and chooses a sequence of consumption,

[13] Auclert et al. (2021)'s algorithm efficiently solves such systems by representing them as DAGs mapping $\mathbf{z}$ *and* $\mathbf{y}$ to 'targets', including the right hand side of the equilibrium conditions themselves (which must equal zero). Their use of DAGs to solve the system does not contradict the fact that the system is nonrecursive.

[14] Simon (1953); LeRoy (2020), describe ways to define the causal effect of y_j on y_i without directional equations, in models with a block-recursive structure. This approach avoids the problems outlined below; but in generic equilibrium models, which are not block-recursive, it will simply say that endogenous variables do *not* causally affect each other.

[15] For now I assume these are equations, not correspondences. I return to this issue in Section 3.5.

[16] Farhi and Werning (2019); Werning (2022) adopt a similar approach. Bollen (1987); Heckman and Pinto (2023) describe how mediation analysis can be applied in simultaneous equation models more generally.

hours worked and real bond holdings $\{c_t, n_t, b_{t+1}\}_{t=0}^{\infty}$, subject to a no-Ponzi condition, to solve

$$\max_{\{c_t,n_t,b_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta_t \left[\frac{c_t^{1-\gamma}}{1-\gamma} - \varphi \frac{n_t^{1+\nu}}{1+\nu} \right]$$
$$\text{s.t. } c_t + \frac{b_{t+1}}{1+r_t} = w_t n_t + T_t + b_t, \qquad b_0 = 0 \text{ given}$$

A continuum of monopolistically competitive firms hire labor to produce differentiated varieties of output using the linear technology $y_t = n_t$ and set prices subject to Rotemberg adjustment costs; competitive final goods firms aggregate these varieties into the consumption good; a central bank sets the nominal interest rate according to a Taylor rule subject to shocks; and (for simplicity) the government runs a balanced budget. This yields the standard three log-linearized equations

$$y_t = y_{t+1} - \frac{1}{\gamma}(i_t - \pi_{t+1}), \qquad \pi_t = \kappa y_t + \beta \pi_{t+1}, \qquad i_t = \phi \pi_t + \varepsilon_t$$

Consider a one-off monetary policy shock: $\varepsilon_0 \neq 0$, $\varepsilon_t = 0$ for $t > 0$. Provided $\phi > 1$, there is a unique bounded solution, in which all variables return to steady state from date 1 onwards: $y_t = \pi_t = i_t = 0$ for $t > 0$. The solution is $y_0 = -\frac{1}{\phi\kappa+\gamma}\varepsilon_0$, $\pi_0 = -\frac{\kappa}{\phi\kappa+\gamma}\varepsilon_0$, $i_0 = \frac{\gamma}{\phi\kappa+\gamma}\varepsilon_0$.

Kaplan et al. (2018) return to the household problem and write date 0 consumption as a function of all the variables the household takes as given, $c_0 = C_0\left(\{r_t, w_t, T_t\}_{t=0}^{\infty}\right)$. Totally differentiating:

$$dc_0 = \underbrace{\sum_{t=0}^{\infty} \frac{\partial c_0}{\partial r_t} dr_t}_{\text{direct effect}} + \underbrace{\sum_{t=0}^{\infty} \frac{\partial c_0}{\partial w_t} dw_t + \sum_{t=0}^{\infty} \frac{\partial c_0}{\partial T_t} dT_t}_{\text{indirect effects}} = -\frac{1}{\gamma} \frac{c_0}{1+r_0} \left(\underbrace{\beta}_{\text{direct effect}} + \underbrace{1-\beta}_{\text{indirect effects}} \right) dr_0$$

where the second equality uses the assumption of a one-time monetary policy shock (see Appendix A.1). The first term reflects the direct effects of a change in interest rates, holding wages and profits constant, due primarily to intertemporal substitution (Kaplan et al., 2018). The second term reflects the indirect effects of changes in wages and profits: lower interest rates directly raise consumption; this increased goods demand raises firm profits and induces firms to increase labor demand, pushing up wages and household income; households respond by further increasing consumption. With $\beta \approx 1$, according to this decomposition, in RANK monetary policy mostly affects output via direct effects, aka intertemporal substitution or the real rate channel.

Having defined the decomposition, we now explore whether it helps us understand the model.

3.2 Which parameters matter?

Often we want to know which parameters are the most important determinants of the response of y to z in our model. We may wish to know which parameter configurations (if any) allow the model to reproduce empirical estimates of this response. Other times, we lack such estimates and must predict the effect of z on y, so we want to know which parameters must be precisely identified in order to do so correctly. Since we cannot solve the model for every possible parameter configuration,

and would not be able to comprehend the results even if we had them, we rely on 'understanding' the model to anticipate which parameters are key. 'Toy' models, approximations, and similar tools can help understand more complex models. For example, Campbell (1994)'s approximate analytical solution to the canonical RBC model suggests this class of models requires implausibly elastic labor supply and transitory productivity shocks to generate recessions without productivity declines. Do direct-indirect decompositions provide similar understanding – in our example, do they help predict which parameters determine the effect of monetary policy on consumption?

One might think that when direct effects (the 'intertemporal substitution channel') account for most of the response of consumption to the shock, the intertemporal elasticity of substitution (IES) γ^{-1} should be among the most important determinants of this response, while other parameters are less important. While this sounds reasonable, it is not true in general. I now describe two slight modifications to the NK model in which 'direct effects' account for almost all of the response to a monetary policy shock, but the IES does not in any way affect the magnitude of this response.

First, as in Holden (2023), suppose the central bank observes the yield on real bonds r_t (e.g. US TIPS), and sets the nominal rate according to the 'real rate rule' $i_t = r_t + \phi\pi_t + \varepsilon_t$.[17] Using the Fisher equation $i_t = r_t + \pi_{t+1}$, our system has solution $\pi_0 = -\frac{1}{\phi}\varepsilon_0$, $y_0 = -\frac{1}{\kappa\phi}\varepsilon_0$, $i_0 = r_0 = \frac{\gamma}{\kappa\phi}\varepsilon_0$. The share of direct effects is still β. But even when $\beta \to 1$, i.e. monetary policy operates *entirely* through an intertemporal substitution channel according to the Kaplan et al. (2018) decomposition, the aggregate consumption response to a monetary policy shock does not depend *at all* on the IES, only on parameters of the Phillips curve and monetary policy rule.

The second modification keeps the standard Taylor rule unchanged, but changes the model's supply side by assuming decreasing returns ($Y_t = n_t^{\alpha}$), rigid real wages (Blanchard and Galí, 2007), and working capital (Christiano et al., 2005; Ravenna and Walsh, 2006): firms must borrow their wage bill $w_t n_t$ from intermediaries at gross nominal interest rate $1 + i_t$, so their real total costs become $wn_t(1 + i_t)$ rather than $w_t n_t$, and the log-linearized Phillips curve is $\pi_t = \kappa y_t + \frac{\kappa\alpha}{1-\alpha} i_t + \beta\pi_{t+1}$. Appendix A.2 shows that the response to a one-period monetary policy shock is $y_0 = -\left[\phi\kappa + \gamma\left(1 - \frac{\phi\kappa\alpha}{1-\alpha}\right)\right]^{-1}\varepsilon_0$. When $\frac{\phi\kappa\alpha}{1-\alpha} = 1$, we have $y_0 = -\frac{1}{\phi\kappa}\varepsilon_0$ as with the real rate rule, and again the IES is irrelevant for the equilibrium response of aggregate output to a monetary policy shock – even when the share of direct effects $\beta \approx 1$. Worse still, when $\frac{\phi\kappa\alpha}{1-\alpha} > 1$, a higher IES (lower γ) *reduces* the sensitivity of output to a given monetary policy shock.[18]

Alternatively, introduce a fraction η of hand to mouth households (HtMs) and suppose the government adjusts lump-sum taxes to keep the real value of nominal debt B_{t+1}/P_t constant. This introduces a new channel: contractionary monetary policy reduces inflation, increasing the real value of outstanding debt, and requiring higher taxes, which reduces spending, especially for HtMs. The share of the aggregate consumption response to ε_0 'accounted for' by these indirect effects

[17] One might feel uneasy about this equation. How can the central bank observe the yield on real bonds before setting the nominal rate, when the former yield is determined by arbitrage between real and nominal bonds? Such questions are reasonable, yet hard to answer in equilibrium models; see Section 4.2 for analysis of a similar model.

[18] While Kaplan et al. (2018) define the direct effect of monetary policy in terms of the change in r_t, they note one could define it 'even more directly' in terms of the shock ε_t. Appendix A.3 shows this does not resolve the paradox: the share of direct effects can still be arbitrarily close to 100% even though the IES is irrelevant for aggregate outcomes.

due to taxes can be substantial (unlike in our baseline RANK economy). One might think this large indirect share indicates that fiscal variables, e.g. the debt-to-GDP ratio $\frac{B}{PY}$, are important determinants of the aggregate response. In fact, Appendix A.2 shows that with a real rate rule, $\frac{B}{PY}$ is irrelevant for the aggregate response, however large the share of indirect effects due to taxes.

To understand why the decompositions 'fail', consider a more general system with an outcome variable y, a $(n-1)\times 1$ vector of other endogenous 'mediators' $\mathbf{x}$, and an exogenous scalar z:

$$\begin{aligned} \mathbf{A}\mathbf{x} + \mathbf{b}y &= \mathbf{c}z \\ y &= \mathbf{g}^{\intercal}\mathbf{x} \end{aligned} \tag{5}$$

where the last block is a behavioral equation relating an agent's choice of y to the variables $\mathbf{x}$ she takes as given. In a recursive system, $\mathbf{b} = \mathbf{0}$ and $\mathbf{A}$ is lower triangular with ones on the diagonal; in an equilibrium model, $\mathbf{A}$ and $\mathbf{b}$ are unrestricted. As in Kaplan et al. (2018), we can decompose the total effect $\frac{dy}{dz} = \frac{dy}{dz}\sum_i s_i$, where $s_i := g_i \frac{dx_i/dz}{dy_i/dz}$ is the share of dy/dz accounted for by x_i. Does s_i help us predict how dy/dz would change if we changed g_i, the direct sensitivity of y to x_i?

Proposition 3.1. *If (5) is recursive, the system has a unique solution and s_i equals the signed elasticity of dy/dz with respect to g_i, $\epsilon_i = \frac{d^2y}{dg_i dz}\frac{g_i}{dy/dz}$. If (5) is unrestricted: (i) it has a unique solution if $\mathbf{A}+\mathbf{b}\mathbf{g}^{\intercal}$ is invertible; (ii) s_i is not a sufficient statistic for ϵ_i; the two may have opposite sign. In particular, if $\mathbf{b}$ is not contained in the column space of $\mathbf{A}$, $\epsilon_i = 0$ whatever the value of s_i.*

Proof. See Appendix A.4. □

In a process model, decomposing the effect of z on y into its effect via the mediators $\mathbf{x} = x_1, ..., x_{n-1}$ *does* help predict how the parameters $\mathbf{g}$ influence this total effect. This is because changes in $\mathbf{g}$, which govern how the agent's choice of y depends on $\mathbf{x}$, do not affect the response of the mediators $\mathbf{x}$ to the shock ($\frac{d^2x_j}{dg_i dz} = 0$): by Principle 1, $\mathbf{x}$ is determined before the agent chooses y. Thus, a change in g_i only affects the response of y, and it does so to the extent that x_i accounts for a large share of the total effect dy/dz. If e.g. dy/dz is entirely accounted for by the agent's response to x_1 (real interest rates r_0), doubling g_1 (the direct effect of r_0 on y) must double dy/dz.

In an equilibrium model violating Principle 1, $\mathbf{x}$ is not 'predetermined': all equations simultaneously determine all variables, and changes in $\mathbf{g}$ can change the response of $\mathbf{x}$ to the shock ($\frac{d^2x_j}{dg_i dz} \neq 0$). The 'shares' s_i are not informative about the effect of g_i on these responses, and so cannot be a sufficient statistic for the effect of g_i on dy/dz: $\frac{d^2y}{dg_i dz} = \frac{dx_i}{dz} + \sum_j g_j \frac{d^2x_j}{dg_i dz}$, and the latter sum is not zero. Equally, while the behavioral equation *can* affect $\mathbf{x}$, it has no special role in determining y.[19] In fact, in our examples, y is completely determined by the rest of the system *excluding the behavioral equation*. One can solve for the consumption response without knowing the IES or the equilibrium value of r_0, even though the 'direct effect' of interest rates accounts for this entire response; the IES only affects what value of r_0 is required to 'implement' this response.

These are deliberately extreme cases: in general, all equations and parameters influence the effect of any shock on any variable. But it remains true that direct and indirect effect shares

[19] In other words, however we represent the model as a causal graph, the graph has cycles.

do not reliably indicate which parameters influence the total effect, and how. To think that the behavioral equation describes direct and indirect 'effects' on y is misleading, and exaggerates the relative importance of *this* equation's parameters; it is just one among many that must be satisfied.

This helps explain why different researchers reach contradictory conclusions about the monetary transmission mechanism in the same RANK models:

> ...monetary policy in RANK models works almost exclusively through intertemporal substitution: direct effects account for nearly the entire impact of interest rate changes on the macroeconomy and indirect effects are negligible. (Kaplan et al., 2018)
>
> ...the transmission mechanism of monetary policy in New-Keynesian models does *not* operate through the real interest rate channel...inflation is approximately determined as in a flexible-price model; output is then pinned down by the New-Keynesian Phillips curve (Rupert and Sustek, 2019)

Such disputes – indeed all attempts to evaluate causal stories told about equilibrium models – are bound to be inconclusive. It is simply meaningless to ask through what 'channels' monetary policy works in a model where all endogenous variables are simultaneously determined.

3.3 Generalization

Interpreting z as an intervention, identifying the mechanisms through which z produces its effects can also help generalize results from one setting to another: knowing *how* something works helps predict *when* and *where* it will work. Pawson and Tilley (1997) (pp78-80) consider a researcher studying how installing closed-circuit television (CCTV) in car parks reduces crime. CCTV might work by facilitating real-time deployment of security staff to areas where crime is occurring; by deterring potential offenders who don't want to get caught on camera; by increasing usage of car parks, so other drivers engage in 'natural surveillance' which deters offenders; etc. Identifying the mechanism helps predict whether CCTV will reduce crime in any *particular* car park, given local context. If CCTV works by facilitating deployment of security staff, it won't work in an isolated car park with no security presence; cars parked in CCTV blind spots will remain vulnerable if the mechanism is deterrence, but not if it is 'natural surveillance'; and so on. In particular, we often wish to generalize to environments where a particular mechanism would be 'damped' or inoperative (e.g. 'facilitating deployment' is inoperative if there are no security guards to deploy).[20]

Similar arguments can be made in economics. For example, if the effect of interest rates on household mortgage payments is an important part of the monetary transmission mechanism (Di Maggio et al., 2017), one might expect policy to be more effective in countries where adjustable rate mortgages are prevalent, but less effective in countries with fixed rate mortgages, where this channel is 'damped'. Similarly, in our simple example, suppose the nominal rate faced by households is

[20] See also Darden (2013) on mechanisms in biology: "One use is to make predictions...if a portion of the mechanism is broken, then one can predict that the earlier stages operated and an intermediate product accumulates (or perhaps no product at all is produced)...A scientist may be able to run a mental simulation of the mechanism and thereby predict what phenomenon it will produce or to predict what will happen if a part of the mechanism is broken."

$(1+i_t)^\lambda R^{1-\lambda}$, where $R=\beta^{-1}$ is the steady state interest rate and $\lambda \in [0,1]$. We know the effect of a monetary policy shock on output in an economy with $\lambda = 1$; a fraction $\beta \approx 1$ of this effect comes via direct effects. We wish to predict how the effect will differ in an otherwise identical economy where the interest rates faced by households react less strongly to the policy rate ($\lambda < 1$).

Since monetary policy operates primarily through a 'direct' real rate channel, one might expect it to be less effective if the channel is damped ($\lambda < 1$). But in our two examples (Holden (2023)'s real rate rule and the working capital economy with $\frac{\phi\kappa\alpha}{1-\alpha} = 1$), $dy/d\varepsilon_0$ does not depend on λ at all. This follows from our earlier results: 'damping' is isomorphic to reducing γ^{-1} to $\lambda\gamma^{-1}$, and γ^{-1} is irrelevant for $dy_0/d\varepsilon_0$. Consumption is pinned down by the rest of the model excluding the household consumption function: if the direct effect of interest rates is muted, rates will endogenously respond more strongly, for the same shock ε_0, to 'implement' the same consumption response.[21]

In our general framework (5), suppose we know dy/dz and wish to generalize to a setting where the response of y to x_i is damped (g_i is replaced with λg_i). In recursive models, the decomposition helps do this – by Proposition 3.1, s_i equals the elasticity of dy/dz with respect to λ – because damping y's response to x_i does not affect the response of the mediators $\mathbf{x}$, which are determined before y is chosen, to z. The behavioral equation is 'directly responsible' for y: a weaker g_i makes y less responsive to z through the x_i channel, since x_i cannot adjust to compensate. But in equilibrium models, all equations determine all variables, so damping g_i *does* affect $d\mathbf{x}/dz$; the shares s_i are uninformative about this effect, and cannot tell us how λ affects dy/dz. The behavioral equation has no special responsibility for y: if g_i is weakened, $\mathbf{x}$ may adjust to produce the same, or even a stronger response of y to z. Since the decompositions do not actually identify *how* z determines y, they do not reliably help us generalize this effect from one context to another.

3.4 Designing alternative interventions

Another reason to study the mechanisms through which an intervention produces its effect is that this helps design better interventions (discarding elements which are not important to the outcome), or substitutes in case the original 'treatment' is not feasible. VanderWeele (2015) discusses a cognitive behavioral therapy (CBT) intervention which reduced depression symptoms, but also increased participants' use of antidepressants. If the intervention worked only by encouraging antidepressant use, directly prescribing antidepressants might be more cost-effective; if it changed participants' thought and behavior patterns in other ways, its CBT aspects should not be discarded.[22] For an economic example, take the debate over the mechanisms through which quantitative easing (QE) – central bank purchases of long-term bonds funded by reserve creation – affects the economy (Carlson et al., 2020). If QE works by changing the amount of duration risk held by private investors, one might achieve the same effects by buying long-term bonds and selling short-term bonds, without

[21] Relatedly, one might expect paradoxes arising in rational expectations models to be mitigated if we damp 'GE effects' by making expectations less sensitive to realized variables. But Section 6 shows that the opposite can happen.

[22] Pearl and Mackenzie (2018) discuss a cautionary tale highlighting how failing to identify the mechanism can lead to poor choice of substitute treatments. Believing citrus fruits prevented scurvy because of their acidity, in the 1860s the Royal Navy replaced Malta lemons with West Indian limes (highly acidic, but containing barely one-quarter as much vitamin C). The British Arctic Expedition of 1875 suffered badly from scurvy despite taking lime juice.

expanding the central bank balance sheet, as in the Fed's 2011 Operation Twist. Conversely, if QE works mainly by increasing reserves and thereby increasing bank lending, Operation Twist would be less effective, and balance sheet expansion might be a necessary side-effect of effective QE.

In our examples, if monetary policy works primarily through intertemporal substitution, what does that suggest about the design of alternative policies to achieve the same effect? Many economists have noted that a temporary reduction in (or commitment to increase) consumption taxes has the same effect on households' incentive to intertemporally reallocate spending as lower interest rates.[23] So one might expect such a tax cut to provide the same stimulus to aggregate consumption as an expansionary monetary policy shock $\varepsilon_0 < 0$. But the tax cut has *no* effect in our two examples: it is completely offset by an endogenous increase in interest rates.

This point is not unique to RANK or to models with a policy rule; it applies whenever behavior depends partly on endogenous variables outside the policymaker's direct control. Consider a stylized spatial search and matching model. Workers search for jobs either in the 'core' where they live, or the 'periphery'. The probability of finding a job in the periphery is $q(\theta) = \sqrt{\theta}$, where $\theta = v/u$ is the ratio of vacancies to searchers in the periphery. The expected utility from searching there is $q(\theta)\delta_w^{-1}\varepsilon$, where $\delta_w > 1$ is the worker's cost of commuting and ε a Pareto distributed taste shock, $Pr(\varepsilon > x) = x^{-\alpha}$, $\alpha > 0$. Normalizing the expected utility of searching in the core (assumed to be fixed) to 1, the fraction of workers searching in the periphery is $u = (q(\theta)/\delta_w)^{\alpha}$. Employers are modeled symmetrically, with commuting cost $\delta_e > 1$, and can post vacancies in either location; a fraction $v = (p(\theta)/\delta_e)^{\alpha}$ posts in the periphery, where $p(\theta) = 1/\sqrt{\theta}$ is the job filling rate.[24]

Starting from an equilibrium where $\delta_w = \delta_e = \delta_0$, suppose a government seeking to increase job creation in the periphery subsidizes transport to the periphery for everyone, reducing commuting costs to $\delta_e = \delta_w = \delta_1 < \delta_0$. This will increase job search and vacancies in the periphery in equal measure, $\frac{d\ln u}{d\ln \delta} = \frac{d\ln v}{d\ln \delta} = -\alpha$, leaving θ unchanged. An economist seeking to understand the mechanisms though which the policy encouraged vacancy posting might decompose its effects:

$$d\ln v = \underbrace{\alpha d\ln p(\theta)}_{\text{worker availability effect}=0} - \underbrace{\alpha d\ln \delta_e}_{\text{commuting cost effect}}$$

This decomposition suggests the subsidy influenced vacancy posting entirely by reducing employers' commuting costs, and not at all by changing worker availability. Thus, our economist might suggest a more targeted policy: a commuting subsidy available only to *employers*, which reduces δ_e to $\delta_1 < \delta_0$ but keeps $\delta_w = \delta_0$. Since the original policy influenced job creation entirely through employers' commuting costs, this targeted policy should (the economist reasons) have the same effect at a smaller fiscal cost. But in fact, this policy is less effective, $\frac{d\ln v}{d\ln \delta_e} = -\alpha + \frac{\alpha^2}{2(1+\alpha)} > -\alpha = \frac{d\ln v}{d\ln \delta}$. While this policy is designed to 'directly' increase employers' vacancy posting, by doing so, it also changes the endogenous variables affecting their decision, increasing market tightness θ (since workers are

[23] Feldstein (2002) suggested that Japan commit to gradually increase sales taxes following its 'lost decade': "This tax-induced inflation would give households an incentive to spend sooner...an expansionary fiscal policy based on a revenue-neutral structural incentive may be more productive and less risky than an excessively easy monetary policy".

[24] If an employer creates a job in the periphery, she must travel there to supervise production.

reluctant to search in the periphery without a subsidy to *their* commuting costs) and reducing the job filling rate, which tempers employers' willingness to take advantage of their subsidy.

In recursive models, direct-indirect effect decompositions can help design treatments by indicating how the factors affecting an outcome y might be independently varied to control y. But in equilibrium models, intervening 'directly' on any one equation changes the value of all endogenous variables in the system; single-equation decompositions cannot reliably predict how the whole system changes. Formally, change (5) to $y = \mathbf{g}^\intercal \mathbf{x} + h\widetilde{z}$ where $h = g_i \frac{dx_i}{dz}$ for some i. $\widetilde{z}$ represents a 'substitute' treatment (e.g. the consumption tax cut) which 'mimics' the effect of z on y via x_i (e.g. the effect of real interest rates on consumption), but does not directly affect the rest of the system. s_i is informative about the effect of w in recursive models, but not equilibrium models.

Proposition 3.2. *If (5) is recursive, the relative effectiveness of $\widetilde{z}$ and z, $\frac{dy/d\widetilde{z}}{dy/dz}$, equals s_i. If (5) is unrestricted and has a unique solution, s_i is not a sufficient statistic for $\frac{dy/d\widetilde{z}}{dy/dz}$; the two may have opposite sign. If $\mathbf{b}$ is not contained in the column space of $\mathbf{A}$, $dy/d\widetilde{z} = 0$ whatever the value of s_i.*

Proof. See Appendix A.5. □

In equilibrium models, even an intervention $\widetilde{z}$ designed to 'directly' affect y without being mediated by $\mathbf{x}$ does in fact affect $\mathbf{x}$, which is determined jointly with y. Not so in process models, where the variables influencing an agent's action must, logically, be determined *before* she acts.

3.5 Some clarifications

I am not claiming that anyone would actually be misled by the decompositions in the above examples, which are deliberately simple. The attraction of decompositions is that they can be applied even in analytically intractable models; the danger is that they could suggest misleading predictions in such cases, where errors are harder to detect. This is a risk to be aware of: it does not imply decompositions are misleading in any particular case; and even if they can be, they may be useful for other purposes. Finally, I do not claim there is *no* way to understand equilibrium models; this paper attempts to do just that. But since attempts to do so using decompositions (and 'causal process' stories generally) are widespread, it is important to understand the limits of this approach.

In the cases above, decompositions can be misleading; but sometimes, they cannot even be performed. Some variables – prices in a competitive model, market tightness in a search model – are not chosen by any agent, and have no corresponding behavioral equation; there is no obvious way to decompose their response into direct and indirect effects. Even when a variable *is* chosen by some agent, their optimization problem may have no solutions, or many. In our NK example, if $\gamma = 0$, (linear utility), the household block reduces to $i_t = \pi_{t+1}$: it pins down the real interest rate but leaves consumption undetermined.[25] In such cases, decompositions are impossible, since there is no 'consumption function' to differentiate. Again, the behavioral condition describing an agent's choice of y has no special role in pinning down y: all conditions determine all variables.

[25] This is far from a rare occurrence in equilibrium models: it can arise from linear disutility of labor supply, no-arbitrage conditions, constant-returns-to-scale production functions, free entry conditions, etc.

The fact that equilibrium models cannot consistently be understood in terms of causal process is a significant disadvantage. While other tools can aid understanding, they are limited. 'Toy' models have a small number of endogenous variables; supply-demand graphs, just two; but causal diagrams are intelligible even in large systems. The intelligibility of causal stories surely explains why economists are drawn to describe their models in these terms. More fundamentally, identifying the mechanisms through which causes produce their effects in the real world is a core goal of science. If equilibrium models do not represent this causal process, they are incomplete at best. At worst, they may generate predictions inconsistent with any plausible causal process, as I now describe.

4 Paradoxes of temporal nonlocality

This section discusses extensive form games exhibiting the following paradox. The subgame following some history has multiple Nash equilibria $\{y_1, y_2, ...\}$. But in an equilibrium of the full game, this history is reached and only one value of y can realize. This is paradoxical: it is intuitively unclear how the prior history determines outcomes in the full game, if not by changing preferences, feasible actions, etc. in the folllowing subgame so as to select a unique subgame equilibrium.

Before observing this paradox in more interesting settings, we examine the simplest possible example, which we call the ***ij* game**. There are two periods, $t, t+1$: a unit continuum of players $i \in [0,1]$ each choose $x_i \in \mathbb{R}$ at date t, and a separate continuum $j \in [0,1]$ choose $y_j \in \mathbb{R}$ at date $t+1$. Each i has payoff $u_i(x_i; x, y) = yx_i$, while each j has payoff $u_j(y_j; x, y) = 0$, where we define the aggregate actions $x = \int x_i di$, $y = \int y_j dj$, (Intuitively, j is always indifferent, while i can make unbounded profits if $y \neq 0$.) An equilibrium is a collection $\{\{x_i\}_{i\in[0,1]}, \{y_j\}_{j\in[0,1]}, x, y\}$ such that[26]

(i-Opt) Each agent $i \in [0,1]$ chooses x_i optimally given x, y: $x_i \in \arg\max_{\widetilde{x}_i \in \mathbb{R}} u_i(\widetilde{x}_i; x, y)$

(j-Opt) Each agent $j \in [0,1]$ chooses y_j optimally given x, y,: $y_j \in \arg\max_{\widetilde{y}_j \in \mathbb{R}} u_j(\widetilde{y}_j; x, y)$

(AA) Aggregate actions are consistent with individual actions: $x = \int x_i di$, $y = \int y_j dj$.

Solving the model, we find that $\{\{x_i\}_{i\in[0,1]}, \{y_j\}_{j\in[0,1]}, x, y\}$ is an equilibrium if and only if $\int y_j dj = y = 0$. Why must we have $y = 0$? Since each agent j is always indifferent among all y_j, assumption j-Opt does not help pin down equilibrium outcomes. Instead, we must use assumption i-Opt. i's optimization problem only has a solution if $y = 0$, in which case he is indifferent (any x_i is optimal). If $y \neq 0$, no x_i is optimal. In our Section 2 terminology, i-Opt can be written as $x_i \in \Gamma_i(y)$ where $\Gamma_i(0) = \mathbb{R}$, $\Gamma_i(y) = \varnothing$ for $y \neq 0$, which violates Principles 1 (date $t+1$ variables determine date t variables) and 2 (Γ_i can be empty-valued). Whether this condition is satisfied by any $(\{x_i\}, \{y_j\})$ – whether i acts optimally at date t – does not depend *at all* on x_i, i's action at date t. It depends *solely* on the average action of the j agents at the date $t+1$. To put this another way, if we *assume* i acts optimally at t, that assumption determines what the js do at $t+1$.

[26] This equilibrium concept assumes that each agent i is atomistic, and does not perceive that his action affects x or the choices of agents j; it also rules out sunspots or randomization devices which allow y_j and y to be stochastic. The first restriction matters (we discuss this below); the second is inconsequential.

Now consider the date $t+1$ subgame following any history $\{x_i\}$. An equilibrium of that subgame is a collection $\{\{y_j\}_{j\in[0,1]}, y\}$ satisfying $y = \int y_j dj$ and j-Opt, given $\{x_i\}$ and x. Given any history, *any* $\{y_j\}_{j\in[0,1]}$ is a subgame equilibrium. Nothing ensures $y = 0$. Hence the paradox: if $y \neq 0$ is possible in the subgame following any history, how does i's optimal behavior at t prevent such outcomes in an equilibrium of the full game? Alternatively, consider the full game, but replace i-Opt with the assumption that x_i is exogenously fixed (e.g. $x_i = 0, \forall i$), rather than chosen optimally by i. In this fixed-x game, any y is an equilibrium. But why should it make a difference to js' date $t+1$ behavior whether $x_i = 0$ was chosen optimally, or exogenously fixed at that level, at date t?

To simplify exposition, this game featured many extreme properties: js are indifferent between any action, i's action is not uniquely determined in equilibrium, individual decision problems may not have solutions, and each agent acts only once. Section 4.3 shows these can all be relaxed while retaining the paradox. First, we study how it can arise in two economically interesting examples.[27]

4.1 The pricing game

This example shows how the paradox matters for a seminal question in macroeconomics: does price flexibility ensure full employment, and if so how?

Time is discrete, $t = 0, 1, ..., T$. A continuum of yeoman farmers (Woodford, 1996) indexed by $i \in [0, 1]$ each produces a different variety of the consumption good. Their preferences are

$$\sum_{t=0}^{T} \beta^t \left[\frac{c_t(i)^{1-\sigma}}{1-\sigma} - v(y_t(i)) \right] \tag{6}$$

where $v', v'' > 0, \sigma \geq 0$, and $c_t(i)$ is a CES aggregate of i's' consumption of each variety, $c_t(i) = \left(\int_0^1 c_t(i,j)^{\frac{\gamma-1}{\gamma}} dj \right)^{\frac{\gamma}{\gamma-1}}$, $\gamma > 1$, and $c_t(i,j)$ denotes i's consumption of the jth variety at date t.

Each period is divided into a 'morning' and 'afternoon'. In the morning, each farmer i posts a price $p_t(i)$ for her variety, denominated in an abstract unit of account. In the afternoon, i observes the profile of prices $\{p_t(j)\}_{j\in[0,1]}$ posted by all farmers, and communicates her desired consumption $c_t(i,j)$ to each seller. Each farmer j is required to produce enough to meet demand: $y_t(j) = \int_0^1 c_t(i,j) di$. Goods are purchased on credit, with no credit limit. Each farmer starts with a zero credit balance, $m_0(i) = 0$, and unspent balances earn a zero nominal interest rate overnight:

$$m_{t+1}(i) = m_t(i) + p_t(i) y_t(i) - \int_0^1 p_t(j) c_t(i,j) dj$$

Following the market games literature (Dubey, 1982), a farmer with a negative balance at the end of date T, $m_{T+1}(i) < 0$, incurs a bankruptcy penalty, modeled as a large utility cost χ; we assume throughout that χ is large enough that households will always choose $m_{T+1}(i) \geq 0$ if possible.

Optimal consumption satisfies $c_t(i,j) = c_t(i) \left(p_t(j)/p_t \right)^{-\gamma}$, where $p_t = \left(\int_0^1 p_t(i)^{1-\gamma} di \right)^{\frac{1}{1-\gamma}}$.

[27] All examples in this Section are explicit games, rather than market-clearing models; as in e.g. Bassetto (2002), this makes it more transparent how paradoxes arise. Similar paradoxes will arise in equilibrium models more generally.

Thus, i faces demand $y_t(i) = y_t\left(p_t(i)/p_t\right)^{-\gamma}$, and we can write her constraints as

$$m_{t+1}(i) = m_t(i) + p_t(i)y_t\left(\frac{p_t(i)}{p_t}\right)^{-\gamma} - p_t c_t(i),\ t = 0, ..., T, \quad m_0(i) = 0,\ m_{T+1}(i) \geq 0 \tag{7}$$

An **equilibrium** of the full game is a collection of aggregate and idiosyncratic prices and allocations $\{p_t, y_t, c_t, \{p_t(i), c_t(i), y_t(i), m_{t+1}(i)\}_{i\in[0,1]}\}_{t=0}^T$ such that (i) for each $i \in [0,1]$, $\{p_t(i), c_t(i), y_t(i), m_{t+1}(i)\}_{t=0}^T$ maximizes (6) subject to (7); (ii) $p_t = \left(\int_0^1 p_t(i)^{1-\gamma}di\right)^{\frac{1}{1-\gamma}}$, and $y_t = c_t = \int_0^1 c_t(i)di$.

Proposition 4.1. *Any equilibrium is symmetric, $c_t(i) = y_t(i) = c_t = y_t$, $p_t(i) = p_t$, $m_{t+1}(i) = 0$ for all $i \in [0,1]$, $t = 0, ..., T$. In any equilibrium, for all t, $c_t = y_t = y^*$ where $1 = \frac{\gamma}{\gamma-1}(y^*)^\sigma v'(y^*)$, and $\beta\frac{p_t}{p_{t+1}} = 1$. The date 0 price level is indeterminate: any $p_0 > 0$ is an equilibrium.*

Proof. See Appendix A.6. □

While the overall price level is indeterminate, the classical dichotomy holds and real outcomes are the same in any equilibrium: the rate of deflation (and hence the real return on money) depends on households' discount factor β, and output is constant and equal to y^*, its 'full employment' level. As in the New Keynesian model, this level is inefficiently low, since households' monopoly power leads them to set prices as a markup over their marginal disutility of labor.

Date T afternoon subgame Rather than studying the whole game, we can also study the subgame beginning in the morning or afternoon of any period. Consider the final subgame beginning in the afternoon of date T, and suppose that (as is the case in full equilibrium) all farmers set the same prices in the morning ($p_T(i) = p_T$) and have the same cash balances ($m_T(i) = 0$). The only choice each farmer makes in this subgame is their consumption decision; this determines each farmer's production, final credit balances are computed, and utility costs of bankruptcy (if any) are incurred. An equilibrium of this subgame is a collection $\{y_t, c_t, \{c_t(i)\}_{i\in[0,1]}\}$ such that (i) for each i, $c_t(i)$ maximizes $\frac{c_t(i)^{1-\sigma}}{1-\sigma} - v(y_t)$ subject to $p_T c_t(i) = p_T y_t$, and (ii) $y_t(i) = y_t = c_t = \int_j c_t(j)dj$.

Any common level of spending and production $c_t(i) = y_t > 0$ is an equilibrium of this subgame. Households plan to spend their entire income, $c_t(i) = y_t$; but their income is someone else's spending. Thus, their spending decisions are strategic complements: a higher average consumption level c_t raises i's income, and hence his optimal consumption, one for one. These complementarities are so strong – in terms of Samuelson (1939)'s Keynesian cross, the MPC is 1 and the consumption function lies on top of the 45 degree line – that any level of aggregate income and spending can be self-fulfilling. Nothing rules out 'recessions' or 'general gluts' in which $y_t < y^*$, or 'booms' in which $y_t > y^*$.[28] Yet prices are not fixed at the 'wrong level': they have the same value as in an equilibrium of the full game, which features full employment. Equilibria with $y < y^*$ are 'effective demand failures' in the sense of Clower (1965); Leijonhufvud (1968, 1973). The fix-price literature following their work (see Drazen (1980); Benassy (1990, 1993) for surveys) began by considering models where prices are fixed at non-market clearing levels; my analysis is closest that of to a few

[28] y^* is not even renegotiation-proof: it is Pareto-dominated by the 'boom' equilibrium $\breve{y}$ satisfying $1 = (\breve{y})^\sigma v'(\breve{y})$.

authors who argue underemployment is possible even at the 'right' prices (Heller and Starr, 1979; Citanna et al., 2001), in particular Roberts (1987, 1989), discussed in Section 4.3.

Date T morning subgame Next, consider the subgame in which agents set prices at the beginning of date T (again assuming $m_T(i) = 0, \forall i$). Using $c_T(i) = \left(\frac{p_T(i)}{p_T}\right)^{1-\gamma} y_T$, optimal price setting uniquely defines i's relative price $p_T(i)/p_T$ as an increasing function of y_T:

$$\left(\frac{p_T(i)}{p_T}\right)^{1+\sigma(\gamma-1)} y_T^{-\sigma} = \frac{\gamma}{\gamma-1} v'\left(y_T\left(\frac{p_T(i)}{p_T}\right)^{-\gamma}\right) \Rightarrow \frac{p_T(i)}{p_T} = f(y_T),\ f' > 0, f(y^*) = 1 \qquad (8)$$

If i anticipates higher demand, and a higher disutility of labor, he raises prices in an attempt to consume more and work less. Thus all farmers must set the same price, $p_T(i) = p_T$, and (8) implies $y_T = y^*$ (while the overall price level p_T is undetermined). The assumption that farmers set prices optimally in the morning pins down afternoon output at y^* – 'price flexibility ensures full employment' – even though the afternoon subgame, given any price level, has multiple equilibria.

Again, we have a paradox. If recessions and booms ($y \neq y^*$) are possible in the afternoon subgame following any profile of prices, how can the assumption that farmers set prices optimally rule out such outcomes in the full game? What happens in the morning that coordinates expectations on $y_t = y^*$ in the afternoon? Alternatively, suppose that in the full date T subgame prices were fixed at some common level $p_T(i) = p_T$ (the same level farmers would have chosen in one equilibrium of the flexible-price game), rather than being freely chosen. With fixed prices, again any y_t is an equilibrium: (8) is no longer an equilibrium condition, and nothing rules out $y_T \neq y^*$. Moving back from fixed to flexible prices eliminates the $y_t \neq y^*$ equilibria, even when farmers choose the same level p_T as in the fixed-price equilibrium. But how can giving farmers the *option* to change prices rule out coordination failures, even when this option is not actually exercised?

To relate this to the ij game, log-linearize (8) and the definition of p_t to yield

$$\widehat{p}_T(i) - \widehat{p}_T = \kappa(\widehat{y}_T - \widehat{y}^*), \qquad \widehat{p}_T = \int_0^1 \widehat{p}_T(i)di \qquad (9)$$

where $\kappa = \frac{\sigma+\varphi}{1-\sigma+\gamma(\sigma+\varphi)} > 0$, $\varphi = \frac{v''(y^*)}{v'(y^*)y^*}$, and hats denote logs. If farmer i expects economic activity to exceed its full employment level, he will set his price above his expectation of the aggregate price level; if he expects weaker demand, he will try to set a lower price than the average. Out of equilibrium, there is nothing contradictory in each farmer fearing weak demand, and setting a price which she expects to be below the average; of course, they cannot all *succeed* in doing so. But the definition of equilibrium requires (violating Principle 1) that all farmers set prices optimally given *correct* expectations of the current price level and *future* output. Whether any symmetric pricing profile $p_T(i) = p_T$ constitutes optimal pricing behavior in the morning does not depend at all on the level of p_T: it depends solely on aggregate output in the afternoon. If $y_T = y^*$, any symmetric price profile is optimal. But if $y_T \neq y^*$, *no* price profile can be optimal – paraphrasing Keynes (1936, p13)'s essentially identical argument, there exists no expedient by which farmers as a whole

can reduce their *relative* price to $f(y_T) < 1$ by posting lower *nominal* prices. Thus, if we *assume* farmers set prices optimally in the morning, that assumption pins down output in the afternoon.

One might argue the 'paradox' only arises because we restricted attention to perfect foresight equilibria. Once we permit sunspot equilibria, (9) implies $\mathbb{E}(\widehat{y}_T - \widehat{y}^*) = 0$: y_T can take any value, provided $\widehat{y}_T - \widehat{y}^*$ is not predictable in the morning. This is (the argument goes) because the afternoon subgame effectively features fixed prices; without monetary policy pinning down aggregate demand, this naturally leads to real indeterminacy. But this argument stretches the definition of fixed prices. Here prices are *posted* – logically implying that the seller first quotes a price, and then the buyer decides how much to purchase at that price – but fully *flexible* (the seller can quote any price). Thus, even if monetary policy could ensure $y_T = y_T^*$ in the afternoon subgame, this would still constitute a failure of the classical dichotomy: price flexibility *alone* does not guarantee full employment. And the sunspot equilibria are arguably even more paradoxical than the perfect foresight equilibrium: agents must coordinate their deviations from y^* to have ex ante mean zero, even though there is no reason, ex post, why they should not always choose $y < y^*$.

Iwai (1981, 2019) discusses a similar model without assuming rational expectations, providing slightly different, but complementary, interpretations of the key equation (9). He notes that rational expectations are impossible whenever $y \neq y^*$: if e.g. $y > y^*$, then whatever firms' average expectation of the aggregate price level, realized prices will necessarily exceed this. Thus, to postulate the rational expectations hypothesis, it is necessary to assume Say's law (i.e. $y = y^*$). My argument is instead that assuming rational expectations *does*, logically, entail $y = y^*$, but this makes no intuitive sense (since $y \neq y^*$ remains an equilibrium of the afternoon subgame).

Date $t < T$ subgames While the date T outcomes just described (equivalently, a one-period model with $T = 0$) illustrate the core of the paradox, similar paradoxes arise in earlier subgames.

Proposition 4.2. *Given equal initial credit balances $m_t(i) = 0$, any equilibrium of the date t morning subgame with equal credit balances is symmetric, with $p_\tau(i) = p_\tau$, $c_\tau(i) = c_\tau = y_\tau = y^*$ for all $i \in [0,1]$ and $\tau = t, t+1, ..., T$, and $p_\tau = \beta p_{\tau-1}$ for all $\tau = t+1, ..., T$. p_t is undetermined.*

Given $m_t(i) = 0$ and uniform prices $p_t(i) = p_t$, any equilibrium of the date t afternoon subgame is symmetric, with $p_\tau(i) = p_\tau$ for all $\tau > t$ and $c_t(i) = c_t = y_t$ for all $\tau \geq t$. Any $y_t > 0$ is an equilibrium, and (if $t < T$) $p_{t+1} = \beta p_t (y_t/y^*)^\sigma$. $y_\tau = y^*$ for all $\tau > t$ and $p_{\tau+1} = \beta p_\tau$ for all $\tau > t$.

Proof. See Appendix A.7. □

In the date t afternoon subgame, y_t is indeterminate. But at any $\tau > t$, we must have $y_\tau = y^*$; otherwise, farmers would not be setting prices optimally in the morning of date τ. For $y_t < y^*$ to be an equilibrium, then, households must anticipate a high real return on money between t and $t+1$, i.e. they must expect the price level to fall to $p_{t+1} = \beta p_t (y_t/y^*)^\sigma < \beta p_t$. Otherwise, they would wish to borrow against higher future income, raising date t spending and production. When $\sigma = 0$, this is particularly stark: any y_t is an equilibrium of the afternoon subgame, but we must have $p_{t+1} = \beta p_t$, regardless of y_t. That is, households' optimal spending behavior at date t does not pin

down date t consumption and employment: it solely pins down the date $t+1$ price level. Yet in the date $t+1$ morning, nothing forces farmers to set prices consistent with these prior expectations.

4.2 The interest-on-reserves game

My second example draws on Hall and Reis (2016), who argue that a central bank can uniquely determine the price level by paying an appropriate interest rate on reserves. Their method exploits no-arbitrage between reserves and real assets: if the real interest rate is r, and the central bank commits to pay $1+r$ units *of real output* for every dollar of reserves, today's price level must be 1, otherwise there would be an arbitrage opportunity. How asset market arbitrage determines goods prices may seem unclear, partly because Hall and Reis (2016)'s competitive equilibrium model does not specify how prices are determined, or what happens if they deviate from target. In the spirit of Bassetto (2002), I recast their model as a market game following Shapley and Shubik (1977).

A unit continuum of traders each have a unit endowment of the single consumption good at dates 1 and 2. All trade occurs at date 1. In the morning of date 1, date 1 goods are traded; in the afternoon, 'real bonds' – claims on date 2 goods – are traded. Each trader must offer for sale all her date 1 endowment in the goods market, and all her date 2 endowment in the bond market; as in Shapley and Shubik (1977), to consume her own endowment, she must go through the market. Traders also hold interest-bearing central bank reserves, with symmetric initial holdings M. To trade, they allocate a portion of their reserves $C_i \geq 0$ and $B_i \geq 0$, respectively, to bid for goods and bonds. Agents can bid $C_i > M$ (intraday credit is allowed); their bond market bids are bounded above, $B_i \leq \overline{B}$ where $\overline{B} > 1$. The central bank also bids $\delta \in (0,1)$ in the bond market (this ensures bond prices are positive). Prices equal the ratio of the sum of nominal bids to the (unit) supply of each good: $p = \int C_i di$ is the current dollar price of date 1 goods, and $q = \int B_i di + \delta$ the date 1 dollar price of date 2 goods. The real interest rate is $1+r = p/q$. A trader bidding (C_i, B_i) secures C_i/p units of the good at date 1, and claims on B_i/q units for delivery at date 1.

After the afternoon market closes, i's reserve balance is $A_i = M + p - C_i + q - B_i$. ($A_i$ may be negative, i.e. the trader may borrow from the central bank.) At date 2, the central bank pays the holder of each dollar of reserves $1+r$ consumption goods, as in Hall and Reis (2016). These goods are acquired by levying a real lump sum tax $T = (1+r)\int A_i di - \delta/q$ on each trader. Thus, a trader's date 2 consumption is $B_i/q + (1+r)A_i - T$.[29] Traders are atomistic and do not internalize the effect of their bids on prices. In the morning of date 1, trader i chooses C_i to maximize

$$\theta \ln\left(\frac{C_i}{p}\right) + \ln\left(\frac{B_i}{q} + (1+r)A_i - T\right) \tag{10}$$
$$\text{s.t. } A_i = M + p - C_i + q - B_i,\ C_i, B_i \geq 0,\ B_i \leq \overline{B}$$

In the afternoon, she chooses A_i, B_i to maximize (10), taking C_i as given. θ is a preference shifter affecting traders' impatience; we assume $(\overline{B}+\delta)^{-1} < \theta < \delta^{-1}$. An equilibrium is a collection

[29] So defined, some strategy profiles imply negative consumption, but these will never be optimal. One can assume agents can produce additional consumption goods if necessary to pay taxes, at an infinite negative utility cost.

$\{\{A_i, B_i, C_i\}_{i\in[0,1]}, p, q, x, T\}$ such that (i) each $\{A_i, B_i, C_i\}$ solves (10) given p, q, x, T; (ii) $p = \int C_i di$, $q = \int B_i di + \delta$; (iii) $1 + r = p/q$; and (iv) $T = (1 + r) \int A_i di - \delta/q$.

In the subgame beginning in the afternoon of date 1, after the price level p has been determined, trader i must allocate her wealth between bonds, which yield $1/q$ goods per dollar invested, and reserves, which yield $1 + r = p/q$ goods per dollar; i.e. she chooses $B_i \in [0, \overline{B}]$ to maximize $[1/q - (p/q)]B_i$. If $p < 1$, the return on bonds is higher, and traders will submit the maximum bid in the bond market, $B_i = \overline{B}$, yielding price $q = \delta + \overline{B}$. If $p > 1$, the return on reserves is higher, and traders will submit zero bids, $B_i = 0$, yielding $q = \delta$. Finally, if $p = 1$, bonds and reserves offer the same return, and traders are indifferent. Thus any $B \in [0, \overline{B}]$ can occur in an equilibrium of the subgame, and the real interest rate $1 + r = 1/q \in [(\overline{B} + \delta)^{-1}, \delta^{-1}]$ is indeterminate.

In the full game beginning the morning of date 1, i's optimal bid C_i satisfies the Euler equation $\theta(C_i)^{-1} = (1+r)\left(\frac{B_i+\delta}{q} + (1 + r)[p - C_i + B - B_i]\right)^{-1}$ which, given i's optimal choice of B_i, implies

$$C_i = \frac{\theta}{1+\theta}\left(p + B + \frac{\delta}{p}\right) + \frac{\theta}{1+\theta}\left(\frac{1-p}{p}\overline{B}\right)^+ \tag{11}$$

where $x^+ := \max(x, 0)$ (see Appendix A.8 for derivations). In equilibrium, $p = C$. If $C < 1$, then $B = \overline{B}$, and (11) implies $C = \sqrt{\theta(\overline{B} + \delta)} > 1$, a contradiction. If $C > 1$, then $B = 0$, and (11) implies $C = \sqrt{\theta\delta} < 1$, a contradiction. So we must have $C = 1$, and (11) implies $B = \theta^{-1} - \delta$.

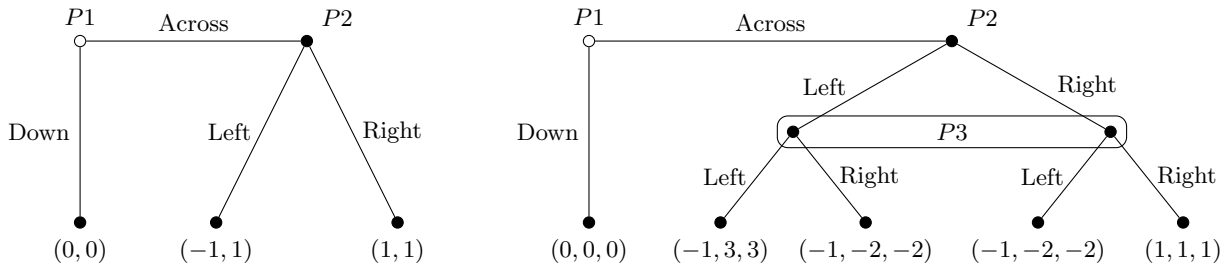
In an equilibrium of the full game, conditions in the morning uniquely determine actions in the afternoon. In particular, changes in households' impatience θ change their bid B in the afternoon. Yet given the prices determined in the morning, and regardless of θ, *any* value of $B \in [0, \overline{B}]$ (hence any interest rate) is an equilibrium of the afternoon subgame, considered as a game in its own right.

4.3 How is this possible?

Each of these paradoxical examples has two essential ingredients. First, it is an extensive form game where the subgame following some history $\mathbf{x}$ has multiple Nash equilibria $Y(\mathbf{x}) := \{\mathbf{y}_1, \mathbf{y}_2, ...\}$. In the pricing game, the subgame following any profile of prices $\{p_i\}$ has multiple equilibrium levels of output; in the interest-on-reserves (IOR) game, the subgame following any history with $p = 1$ has multiple equilibrium levels of nominal bond purchases and interest rates. In general, subgame multiplicity may arise because of strategic complementarities (as in the pricing game, where each farmer wants to spend as much as the average farmer), or because some agents are indifferent (as in the IOR game, where traders are indifferent about their allocation between bonds and reserves).

We might attempt to solve for a subgame perfect Nash equilibrium (SPNE) of such a game by backward induction: For each history $\mathbf{x}$, make *any* desired selection $\mathbf{y}(\mathbf{x})$ from the set of subgame equilibria $Y(\mathbf{x})$. Substitute this into players' payoffs to yield a 'reduced' normal-form game in which players only choose $\mathbf{x}$, and solve for its Nash equilibrium. Such an equilibrium would be guaranteed to exist, had we started from a game with finite actions and players. But our examples have infinitely many actions and players, so there is no such guarantee. In fact, the second key

Figure 3: Finite games



(a) Indifference (b) Coordination

ingredient of the paradox is that this 'reduced game' has no equilibrium for some selections $\mathbf{y}(\mathbf{x})$.[30]

Ingredient 1: subgame multiplicity While both ingredients are necessary for the full paradox, the first ingredient can generate counterintuitive outcomes without the second. Consider a game with subgame multiplicity but finite actions and players, so equilibria of the 'reduced game' exist for any selection $\mathbf{y}(\mathbf{x})$. Agents' optimal choice of $\mathbf{x}$ generally depends on which selection we make: $\mathbf{x}^*$ may only be an equilibrium of the 'reduced game' if agents expect some particular subgame equilibrium $\mathbf{y}^* \in Y(\mathbf{x}^*)$ to be played following $\mathbf{x}^*$. In this case, assuming Nash equilibrium implies that if we reach the $\mathbf{x}^*$ subgame, $\mathbf{y}^*$ must be played, even though this subgame has other equilibria. This is counterintuitive: once $\mathbf{x}^*$ is reached, what stops agents playing another equilibrium?

Figure 3 shows two examples. In the two-player game in the left panel, P1 moves first and chooses $x \in \{\text{Down}, \text{Across}\}$; if she chooses Down, the game ends and each player gets a zero payoff. If she chooses Across, P2 chooses $y \in \{\text{Left}, \text{Right}\}$; either action gives P2 a payoff of 1, but if he chooses Left, P1 gets -1, while if he chooses Right she gets 1. This game has two pure strategy equilibria. If $P2$ finds himself at history $x = \text{Across}$, either Left or Right is optimal: this subgame has multiple equilibria arising from indifference. Suppose he would choose Left at this node, i.e. we make the selection $y(\text{Across}) = \text{Left}$. Anticipating this, $P1$ knows that Across yields a payoff of -1, while Down yields 0, so she will play Down. Suppose instead $P2$ plays Right at $x = \text{Across}$. Anticipating this, P1 will play Across since this yields 1 while Down yields zero. The one outcome that is ruled out is (Across, Left). In other words, $x = \text{Across}$ *can* be reached in equilibrium, but if it is reached, it must be that P2 plays Right. Yet in the subgame starting from $x = \text{Across}$, both $y = \text{Left}$ and $y = \text{Right}$ are valid equilibria.

Figure 3b shows a three-player variant where subgame multiplicity arises from coordination, rather than indifference. Again, P1 moves first. If she plays Down, the game ends with zero payoffs; if she plays Across, P2 and P3 play a coordination game, receiving 3 if they both play Left, 1 if they both play Right, and -2 otherwise. This subgame has two pure strategy equilibria, (Left, Left), giving P1 -1, and (Right, Right), giving P1 1. P1 only plays Right if he expects P2

[30] I focus on nonstochastic selections $\mathbf{y}(\mathbf{x})$. If we permit stochastic selections, there may be additional equilibria where the *distribution* of $\mathbf{y}$ is pinned down, but any $\mathbf{y} \in Y(\mathbf{x})$ can occur with positive probability. The paradox remains: what ensures agents randomize with just the right probabilities, since any $\mathbf{y}$ is always a subgame equilibrium?

and P3 to coordinate on (Right, Right), so (Across, Left, Left) can never arise in a pure strategy equilibrium of the full game. But in the subgame following $x =$ Across, nothing rules out (Left, Left). Indeed, this subgame equilibrium Pareto dominates (Right, Right) for P2 and P3.

These examples lack the full force of the pricing and IOR game paradoxes, in which the assumption that agents optimize in the morning selects a particular equilibrium of the afternoon subgame. Here, the assumption that P1 optimizes ('P1-Opt') does not *determine* how P2 would behave if $x =$ Across were reached: P2 may play either Left or Right in equilibrium.[31] But P1-Opt does impose a particular *relation* between P1's choice of x and P2's behavior following $x =$ Right, violating Principle 1. It is intuitively unclear why this relation should hold. Why can't P1 make a mistake, and play Across, only to find that P2 plays Left? P1 surely cannot *know* ex ante whether P2 will choose Left or Right, since P2 is free to choose either, and both options are rational. Whether P1 behaves optimally is out of her control – the future is undetermined, and any action she takes may prove to be a mistake. Instead, the truth of P1-Opt depends on P2's behavior. To defend this assumption, we would need to explain why *P2* will choose to behave in a way that makes it true.

Ingredient 2: nonexistence Similarly, in the pricing and IOR games, the truth of the assumption that agents optimize in the morning depends on play in the afternoon. But in these games, imposing this assumption does not just imply a *relation* between behavior in the morning and afternoon: it completely pins down afternoon behavior. In the pricing game, farmers must consume y^*, even though any level of consumption is a subgame equilibrium; in the IOR game, traders must bid $B = \theta^{-1} - \delta$ following $p = 1$, even though any bid is optimal. This is because these games feature conditions on payoffs, strategies and players that – while standard – imply that no Nash equilibrium exists in the *reduced game* for some selections $\mathbf{y}(\mathbf{x})$.

In the ij game, the key condition was i's unbounded strategy set. Take any constant selection $y(\mathbf{x}) = \widetilde{y}$ from the equilibria of the $t+1$ subgame; this generates the reduced game where each i maximizes $u_i(x_i; x, y(\mathbf{x})) = \widetilde{y}x_i$. Since i can make unbounded profits for some values of y, an equilibrium of this reduced game is not guaranteed to exist; and indeed it does not, unless $\widetilde{y} = 0$. Unbounded profits is not a pathological assumption: it could arise from free entry, no-arbitrage, constant-returns technologies, etc. Assuming finitely many players, rather than a continuum, would not ensure existence in this case. But if the strategy sets are bounded, say $[-1, 1]$ for all i, j, any $\widetilde{y} \neq 0$ can arise as an equilibrium of the full game, with $x = 1$ if $\widetilde{y} > 0$, and $x = -1$ if $\widetilde{y} < 0$.

Here, non-compact strategy sets imply that i's maximization problem has no solution for any $y \neq 0$ (violating Principle 2), since i can obtain unbounded payoffs. Non-compact strategy sets can lead to nonexistence even when payoffs are bounded and optimization problems have solutions. Consider a two-player caricature of the pricing game. In the afternoon, players $1, 2$ each choose $y_i \in \{0, 1\}$. In the morning, they name integers p_1, p_2. Player 1's payoff is $\mathbf{1}\{y_1 = y_2\} + y_2\mathbf{1}\{p_1 = p_2\} + (1 - y_2)\mathbf{1}\{p_1 < p_2\}$; 2's payoff is symmetric. That is, player 1 receives a dollar if she chooses

[31] Nor does P1's action causally determine P2's future behavior. It is true that P1-Opt implies that if P1 plays Across, then P2 plays Right. But this statement is a material conditional, not a causal conditional: it is true if and only if it is *not* the case that P1 plays Across and P2 plays Left.

the same value of y as 2; also, if 2 chooses $y = 1$, 1 receives an extra dollar if she chose the same integer as 2 in the morning, while if 2 chooses $y = 0$, 1 gets a dollar if she chose a lower integer. In any equilibrium of the whole game, $p_1 = p_2$ and $(y_1, y_2) = (0, 0)$. But whatever integers were announced in the morning, either $(1, 1)$ or $(0, 0)$ is an equilibrium of the afternoon subgame. The only reason $(0, 0)$ cannot occur in full equilibrium is that *if* this had been anticipated, the 'reduced game' played in the morning would become an integer game, which has no equilibrium. Analogously, $y < y^*$ cannot occur in the pricing game. If it did, farmers would each try to undercut the other's price in the morning; this 'undercutting game' has no equilibrium.[32]

Even with compact strategy sets, the second ingredient can arise from a continuum of actions or players (again, standard assumptions). Modify the ij game so each player's strategy set is $[-1, 1]$ and each j has payoff $u_j(y_j; x, y) = -(y_j - y + x)^2$; call this the **discoordination game**. (Rather than being indifferent, j wants to choose a similar action to other $t + 1$ players and a different action than t players.) The equilibrium of the date $t + 1$ subgame is $y_j = 1, \forall j$ if $x < 0$ (each j wants to choose a higher action than the average, leading to a corner solution); and $y_j = -1, \forall j$ if $x > 0$. If $x = 0$, $y_j = y$ is an equilibrium for any $y \in [-1, 1]$. Thus, in any SPNE of the full game, $x = y = 0$. If $x < 0$, $y = 1$, and $x_i = 1$ is optimal for each i, a contradiction; if $x > 0$, $y = -1$, and $x_i = -1$ is optimal, a contradiction. So we must have $x = 0$; but this is only optimal if $y = 0$.

Since any $y \in [-1, 1]$ is an equilibrium following the history $x = 0$, why can't we select $y(0) = \widetilde{y} > 0$, say, and use backward induction to solve the reduced game? This would produce a static game among agents $i \in [0, 1]$ who each have payoff $y(x)x_i$, where $y(x) = 1$ for $x < 0$, $y(0) = \widetilde{y}$, $y(x) = -1$ for $x > 0$. Since this function is discontinuous at $x = 0$, the best response correspondence $\Gamma_i(x) := \arg\max_{x_i} y(x)x_i$ is not guaranteed to be upper hemicontinuous, and in fact it is not: $\Gamma_i(x) = \{1\}$ for $x \leq 0$, $\{-1\}$ for $x > 0$. This reduced game has no equilibrium.

If instead there are n first-period players who each internalize their effect on $x = n^{-1}\sum_{i=1}^{n} x_i$, we *can* construct a SPNE with $x_i = 0, \forall i$ and $y = \widetilde{y} > 0$. Take any i: given that $x_k = 0$ for $k \neq i$, i realizes that $x = x_i/n$, and chooses x_i to maximize $y(x_i/n)x_i$. Choosing $x_i = 0$ yields payoff 0. Any other choice $x_i \neq 0$ triggers an unfavorable action by the js: $x_i < 0$ elicits $y(x) = 1$, yielding payoff $x_i < 0$, while $x_i > 0$ elicits $y(x) = -1$ and payoff $-x_i < 0$. Thus, in the continuum game, the standard 'competitive' assumption of atomistic agents was responsible for the paradox.

The conditions which can generate the paradox – strategic complementarities or indifference to generate subgame multiplicity, atomistic agents or unbounded strategy sets to generate nonexistence – are not pathological, but common in equilibrium models. Ruling out the paradox with restrictions on primitives (finite players and strategies, as in Nash (1950)) would prohibit almost all modern macroeconomic models. We now discuss why nonexistence holds in our economic examples.

Pricing game Consider the static pricing game ($T = 0$), and drop time subscripts. Appendix A.9 shows that following any price profile $\mathbf{p} := \{p(j)\}_{j\in[0,1]}$, any $y > 0$ is an equilibrium, with i's

[32] This is analogous to the use of integer games to rule out undesired equilibria in implementation theory (Jackson, 2001), with two differences. First, here the integer game occurs *before*, not after, the equilibrium it rules out: equilibria exist in every subgame, but not every 'reduced game'. Second, here there is no planner implementing desired outcomes using integer games. Rather, similar 'devices' determine equilibrium outcomes in laissez-faire environments.

corresponding level of consumption given by $c(i) = \left(\frac{p(i)}{p}\right)^{1-\gamma} y$. Take any selection $y > 0$, which depends on the whole profile $\mathbf{p}$: $y(\mathbf{p})$. We continue to assume agents are (and perceive themselves to be) atomistic: $\frac{\partial y(\mathbf{p})}{\partial p(j)} = 0$ for any j. In particular, consider a selection in which $y(\mathbf{p}) \neq y^*$ for every uniform price profile $\mathbf{p}$. We know all farmers choose the same price; but if $y \neq y^*$ following such a price profile, it cannot be optimal to choose $p(i) = p$. Thus, no equilibrium of the pricing game would exist with such a selection. This nonexistence result implies that the selection must feature $y(\mathbf{p}) = y^*$ for some uniform pricing profiles, and one such profile must be played in equilibrium.

Both unboundedness of (log) prices, and atomistic agents, are necessary for this nonexistence result. If each farmer faces bounds on pricing $\underline{p} \leq p(i) \leq \overline{p}$ as in Kocherlakota (2021), her optimality condition implies $p(i) = \max\{\underline{p}, \min\{pf(y), \overline{p}\}\}$. Thus, we can construct an equilibrium in which $y < y^*$ after every uniform pricing profile and $p(i) = \underline{p}$ for all i (or alternatively, $y > y^*$ and $p(i) = \overline{p}$). If farmers expect to sell less than y^*, they each want to undercut their rivals. This 'undercutting game' had no equilibrium when prices were unbounded. With bounds on price setting, it has an equilibrium in which the lower bound binds, so Nash equilibrium does not rule out $y < y^*$.

Next, remove the bounds on pricing but suppose there are n farmers rather than a continuum. As before, they each maximize $\frac{c_i^{1-\sigma}}{1-\sigma} - v(y_i)$; now, i's consumption c_i is a CES aggregate $\left(\sum_{j\neq i} \varepsilon^{-\frac{1}{\gamma}} c_{ij}^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}}$ of varieties produced by other farmers $j \neq i$ (defining $\varepsilon = (n-1)^{-1}$). In the afternoon, i observes all prices $\mathbf{p} := \{p_j\}$ and chooses demand c_{ij} for each variety j, optimally choosing $c_{ij} = \varepsilon(p_i/p_{-i})^{-\gamma} c_i$, where $p_{-i} := \left(\sum_{j\neq i} \varepsilon p_j^{1-\gamma}\right)^{\frac{1}{1-\gamma}}$ is her personal price index; thus, we treat i as choosing c_i directly. Each farmer is required to produce to satisfy demand, $y_i = \sum_{j\neq i} c_{ji}$. Since the economy ends after afternoon trading, it is optimal to spend one's entire income, $c_i = (p_i/p_{-i})y_i$. In the morning, each farmer posts the price p_i for her variety. A SPNE is a pricing profile $\mathbf{p}$ and functions $\mathbf{c}(\mathbf{p}) := (c_1(\mathbf{p}), ..., c_n(\mathbf{p}))$, $\mathbf{y}(\mathbf{p}) := (y_1(\mathbf{p}), ..., y_n(\mathbf{p}))$ such that:

1. for each i, p_i maximizes $\frac{c_i(\mathbf{p})^{1-\sigma}}{1-\sigma} - v(y_i(\mathbf{p}))$ given $\{p_j\}_{j\neq i}$
2. for each i and each $\mathbf{p} \in \mathbb{R}^n_+$, $c_i(\mathbf{p}) = \frac{p_i}{p_{-i}} y_i(\mathbf{p})$ and $y_i(\mathbf{p}) = \sum_{j\neq i} \varepsilon \left(\frac{p_i}{p_{-j}}\right)^{-\gamma} c_j(\mathbf{p})$

As in the continuum game, the afternoon subgame has multiple equilibria for any $\mathbf{p}$: if $(\mathbf{c}(\mathbf{p}), \mathbf{y}(\mathbf{p}))$ is an equilibrium, so is $(\alpha\mathbf{c}(\mathbf{p}), \alpha\mathbf{y}(\mathbf{p}))$ for any $\alpha \in \mathbb{R}$. One possible selection from this set is:

$$c_i(\mathbf{p}) = p_{-i}^{-\gamma} p_i^{1-\gamma} \left(n^{-1} \sum_j p_j^{1-\gamma}\right)^{\frac{2\gamma-1}{1-\gamma}} y^*, \qquad y_i(\mathbf{p}) = p_{-i}^{1-\gamma} p_i^{-\gamma} \left(n^{-1} \sum_j p_j^{1-\gamma}\right)^{\frac{2\gamma-1}{1-\gamma}} y^*$$

where potential output y^* now satisfies $1 = \frac{\widetilde{\gamma}}{\widetilde{\gamma}-1}(y^*)^\sigma v'(y^*)$, as farmers internalize that they face elasticity of demand $\widetilde{\gamma} := \gamma - n^{-1}(2\gamma - 1) < \gamma$, since their pricing decision affects the aggregate price level. Appendix A.9 shows that given this selection, every equilibrium of the full game features uniform pricing, $p_i = p_j$ for all i, j, and $y_i = c_i = y^*$ for all i. Since farmers face less elastic demand than in the continuum model, they set higher markups, leading to lower output; as $n \to \infty$, each farmer's effect on aggregates vanishes, and this equilibrium converges to that of the continuum

game. Again, if we had attempted to modify the selection $(\mathbf{c}(\mathbf{p}), \mathbf{y}(\mathbf{p}))$ above by replacing y^* with $\widetilde{y} \neq y^*$, we would find no such equilibrium of the full game. Thus, even with finite agents, *some* selections from the set of subgame equilibria imply nonexistence of equilibrium in the reduced game.

However, we can make another selection which *does* imply $y < y^*$ on equilibrium. For example:

$$c_i(\mathbf{p})) = p_{-i}^{-\gamma} p_i^{1-\gamma} \left(n^{-1} \sum_j p_j^{1-\gamma} \right)^{\frac{2\gamma-1}{1-\gamma}} \widetilde{y} \prod_j p_j^{\chi}, \qquad y_i(\mathbf{p})) = p_{-i}^{1-\gamma} p_i^{-\gamma} \left(n^{-1} \sum_j p_j^{1-\gamma} \right)^{\frac{2\gamma-1}{1-\gamma}} \widetilde{y} \prod_j p_j^{\chi}$$

where χ satisfies $1 = \frac{\widetilde{\gamma}-\chi}{\widetilde{\gamma}-\chi-1}(\widetilde{y})^{\sigma} v'(\widetilde{y})$. With this selection, $p_i = 1$, $c_i = y_i = \widetilde{y}\ \forall i$ is an equilibrium. For example, we can sustain $\widetilde{y} < y^*$ with $\chi > 0$: each farmer correctly perceives that cutting her price p_i below 1 would reduce every individual's consumption and production by a factor $p_i^{\chi} < 1$, negating the benefit from cutting one's relative price and obtaining a larger share of aggregate output. Further, this effect does not vanish in the limit as $n \to \infty$.[33]

In a similar vein, Roberts (1987, 1989) studies equilibria of a market game with a finite number of agents in which involuntary unemployment occurs at flexible, Walrasian prices. Firms facing zero demand correctly perceive that changing their posted prices and wages would not increase sales; a firm facing positive demand correctly perceives that deviating from competitive prices and wages would trigger a subgame equilibrium in which households 'shun' that firm, neither submitting orders or offering labor. As in my n-farmer game, key to this equilibrium construction is that agents internalize how their individual pricing decisions affect which subgame equilibrium is selected.

IOR game Similarly, we now modify the IOR game by assuming there are n traders who each internalize the effect of their bids (C_i, B_i) on aggregate variables, $p = C = \varepsilon \sum_i C_i$, $B = \varepsilon \sum_i B_i$, $q = B + \delta$ (now defining $\varepsilon = 1/n$). Otherwise, the game is identical to the continuum version (we assume $\theta = 1$ for simplicity). Substituting out for aggregate variables, each trader's payoff becomes

$$U_i(\mathbf{C}, \mathbf{B}) = \ln\left(\frac{C_i}{\varepsilon C_i + C_{-i}}\right) + \ln\left(\frac{B_i + (\varepsilon C_i + C_{-i})[C_{-i} - (1-\varepsilon)C_i + B_{-i} - (1-\varepsilon)B_i] + \delta}{\varepsilon B_i + B_{-i} + \delta}\right) \tag{12}$$

where $C_{-i} = \varepsilon \sum_{j \neq i} C_j$, $B_{-i} = \varepsilon \sum_{j \neq i} B_j$, $\mathbf{C} = (C_1, ..., C_n)$, $\mathbf{B} = (B_1, ..., B_n)$. A SPNE is a profile of goods market bids $\mathbf{C}$ and a function $\mathbf{B}(\mathbf{C})$ such that (1) for each i, C_i maximizes $U_i(\mathbf{C}, \mathbf{B}(\mathbf{C}))$ given $\{C_j\}_{j \neq i}$, and (2) for each $\mathbf{C} \in \mathbb{R}^n$ and each i, $B_i(\mathbf{C})$ maximizes $U_i(\mathbf{C}, \mathbf{B}(\mathbf{C}))$ given $\{B_j(\mathbf{C})\}_{j \neq i}$.

First, consider equilibria in the afternoon subgame following the history $\mathbf{C} = \mathbf{C}^* := (1, 1, ..., 1)$ (in which the price level $p = \sum_i C_i = 1$ attains the central bank's target). Given this profile of bids, any choice of $B_i \in [0, \overline{B}]$ yields i utility 0, and is optimal. With a continuum of atomistic traders, only profiles with $B = \sum_i B_i = 1 - \delta$ were sustainable as an equilibrium of the full game. With a finite number of traders, Appendix A.10 shows that any selection $\mathbf{B}(\mathbf{C}^*)$, together with $\mathbf{C} = \mathbf{C}^*$, is sustainable as an equilibrium of the full game. Pick any $\mathbf{B}$ with $B = \sum_i B_i \in [0, \overline{B}]$, $B \neq 1 - \delta$. To

[33] The selection implies that the afternoon equilibrium becomes infinitely sensitive to the whole profile of price changes: the elasticity to a common change is approximately $n\chi$ which $\to \pm\infty$ as $n \to \infty$. This unappealing property can be avoided: if we wish to sustain $y < y^*$, for example, we can replace the $\prod_j p_i^{\chi}$ term with $\min_j p_j^{\chi}$.

sustain this in equilibrium, we must show that no trader i – say $i = 1$, without loss of generality – prefers to choose $C_1 \neq 1$, if all others choose $C_j = 1$. Thus, we need to compute subgame equilibria following the history $(C_1, 1, ..., 1)$. Following such a history, traders' payoffs become

$$U_1(\mathbf{C}, \mathbf{B}) = \ln\left(\frac{C_1}{\varepsilon C_1 + 1 - \varepsilon}\right) + \ln\left(\frac{B_1 + (\varepsilon C_1 + 1 - \varepsilon)[(1-\varepsilon)(1-C_1) + B_{-1} - (1-\varepsilon)B_1] + \delta}{\varepsilon B_1 + B_{-1} + \delta}\right)$$

$$U_i(\mathbf{C}, \mathbf{B}) = \ln\left(\frac{1}{\varepsilon C_1 + 1 - \varepsilon}\right) + \ln\left(\frac{B_i + (\varepsilon C_1 + 1 - \varepsilon)[\varepsilon(C_1 - 1) + B_{-i} - (1-\varepsilon)B_i] + \delta}{\varepsilon B_i + B_{-i} + \delta}\right) \text{ for } i > 1$$

Appendix A.10 shows that if trader 1 attempts to consume less and save more ($C_1 < 1$), this triggers a fall in prices. Given the Hall and Reis (2016) rule, this raises the real return on bonds above the real return on reserves. All investors invest as much as possible in bonds to exploit the arbitrage, causing interest rates to plummet and making it a bad time to save, reducing 1's utility. Conversely, if 1 tries to consume more ($C_1 > 1$), prices rise, all other traders invest in higher-yielding reserves rather than bonds, interest rates spike, and it is a bad time to borrow, again reducing 1's utility. Thus, even if the selection $\mathbf{B}(\mathbf{C}^*)$ implies that real interest rates r differ from their equilibrium value of 0, this is sustainable in equilibrium: traders recognize that any attempt to adjust their intertemporal spending in response would backfire, causing a discontinuous change in interest rates. As in the pricing game, this remains true in the limit as $n \to \infty$.

4.4 Why does this matter?

Having explained *how* equilibrium assumptions can generate paradoxical outcomes, we now discuss precisely what is paradoxical about these outcomes, and why this should lead us to distrust the assumptions. In our examples, 'history' – features of the environment in the morning – restricts behavior in the afternoon subgame, without affecting any intervening state variables at the beginning of this subgame. In the pricing game, whether y is fixed or indeterminate in the afternoon does not depend on the price *level*, but only on whether prices were freely set or fixed at this level. In the IOR game, traders' preference for early consumption θ affects bids in the afternoon market, but by the time these bids are made, consumption has been determined, θ is irrelevant, and traders are indifferent. This seems to violate 'temporal locality' (TL), the principle that the state of the world at time t can only affect outcomes at $t' > t$ insofar as it affects the state of the world immediately prior to time t'.[34] TL is not logically necessary – it may not even apply to fundamental physical processes (Adlam, 2018) – but for all practical purposes, it accurately describes human behavior. History can only influence behavior today by affecting people's knowledge, beliefs, preferences, opportunities, property rights, etc. today.[35] If equilibrium assumptions really imply that date t interventions affect date $t + 2$ behavior without affecting *anything* about the world at date $t + 1$, this is so contrary to our knowledge of human behavior that we should abandon these assumptions.

[34] See Lange (2002) for a more careful definition.

[35] Davies (2023) goes so far as to argue that "Bygones are bygones" is the one principle that all economists would agree to: "I think it has to be constitutive of what it means to be an economist that you're only going to consider systems and models with a causal structure that respects the one-way flow of time. History can only be brought into an economic model...as a way of talking about an unobserved or unobservable property of the present system."

The only way to reconcile these paradoxical predictions with TL is to suppose that history affects current behavior through 'hidden state variables' not explicitly described by the models. The obvious candidates are agents' mental states: their *knowledge* that there was an earlier period, *memory* of the environment in that period, and *beliefs* about other agents' future conduct. Perhaps we can find a plausible 'belief function' – a process through which agents' memory of the morning subgame and knowledge of the environment causally determine their beliefs about others' behavior in the afternoon subgame – which implies that their behavior in that subgame depends on history exactly as in the paradoxical equilibria. In that event, we would not have to abandon equilibrium assumptions; rather, we would understand more deeply why the assumptions might be correct.

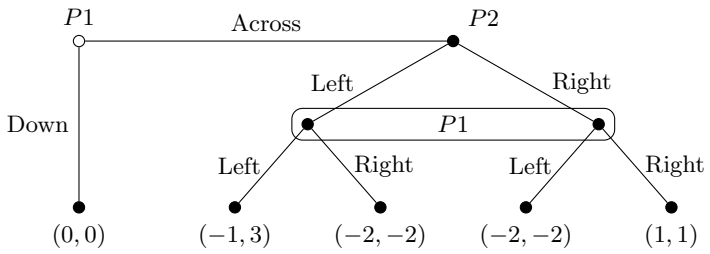
We can only hope to rationalize paradoxical outcomes in this way when subgame multiplicity arises from strategic complementarities. If multiplicity arises from indifference, *no* beliefs about others' behavior would explain why an agent chooses one of the actions in her optimal set rather than another. When multiplicity arises from complementarities, by contrast, we can trivially rationalize *any* equilibrium $(\mathbf{x}^*, \mathbf{y}^*)$ of the multiperiod game $\mathcal{G}$ by assuming that when agent j in the afternoon subgame remembers she is playing $\mathcal{G}$, and observes history $\mathbf{x}^*$, she expects others to play $\mathbf{y}^*$ (and will therefore play y_j^* herself). The question is whether we can find a *plausible* belief function. In the pricing game, we could simply *assume* each farmer expects others to demand y^* if she remembers that prices were freely chosen, but not if she remembers prices were fixed. But this just restates the problem: *why* should the fact that prices were chosen freely cause farmers to expect $y = y^*$?

Since we began with a well-specified game, replacing the assumption of Nash equilibrium with a belief function yields a bona fide process model satisfying the four principles in Section 2. To see whether paradoxical outcomes can be reconciled with TL, then, we are brought back to an old question in economics: whether the *assumption* of equilibrium can be justified by explicitly describing the process through which equilibrium beliefs are acquired. We now discuss whether two such processes studied in the literature, **eductive** and **learning** (or 'evolutive') processes (Binmore, 1987), can rationalize paradoxical equilibrium outcomes in environments like the pricing game.[36]

Eductive justifications The eductive approach describes an internal mental process, proceeding in 'virtual time' in a one-shot game, in which agents try to deduce how others will behave based on beliefs about the game structure and others' rationality. Each agent is rational and only plays strategies that are best responses to *some* probability distribution over other agents' strategies; strategies that are not best responses can be eliminated. But each agent knows others are rational and will eliminate these strategies; acknowledging this, she discovers some of her remaining strategies are no longer best responses, and eliminates them; and so on. This process converges to the set of *rationalizable strategies* (Bernheim, 1984; Pearce, 1984; Guesnerie, 1992). Defining this set in extensive form games is not straightforward, since one must describe how agents update beliefs about others' future actions after observing others' prior actions – in particular, actions inconsistent with rational play. I use Pearce (1984)'s definition of extensive form rationalizability (EFR).

[36] Another process that might coordinate beliefs on one equilibrium is preplay communication. Modeling nonequilibrium communication still requires specifying a learning or eductive 'belief function' (as in e.g. Crawford (2017)).

Figure 4: Modified coordination game



These complications do not arise when each agent only moves once (so no one ever needs to update beliefs about someone's future action after observing their past actions), as in the variants of the ij game in Section 4.3. In these games, EFR does *not* select the paradoxical equilibrium. Any selection $\mathbf{y}(\mathbf{x})$ from the set of subgame equilibria is rationalizable: each j can simply note $\mathbf{x}$, and rationally expect other js to play $\mathbf{y}(\mathbf{x})$. For example, in the discoordination game, $\{-1, 0, 1\}$ are all rationalizable strategies for i. $x_i = -1$ is the best response to the belief that others will play $x = 1, y = -1$ or $x = 0, y \geq 0$; $x_i = 1$ is the best response to $x = -1, y = 1$. Following $x = 0$, j can rationalize any $y \in [-1, 1]$ by the belief that others will play accordingly; nothing singles out the paradoxical equilibrium $y = 0$. True, if j observes $x_i = 0$ for all i, she can infer that all the is expected $y = 0$, but she need not share this expectation herself; maybe the is made a mistake.

When the same agent acts more than once, however, outcomes which would be equilibria in an isolated subgame may not be rationalizable when this subgame arises within a larger game. Figure 4 modifies Figure 3b so that $P1$ both decides whether the subgame following $x =$ Across will be played, and participates in this subgame himself.[37] The full game, as in Figure 3b, has two pure strategy SPNEs, depending on which of the two pure strategy equilibria, (Left, Left) and (Right, Right), we select. However, the only EFR outcome is (Across, Right, Right). Observing P1 play Across in the first round rationally leads P2 to expect P1 to play Right in the subgame. For had P1 expected Left to be played in the subgame, he would have anticipated a payoff of -1 at best, and would have played Down in the first round, avoiding the subgame and receiving 0. Thus (P2 concludes after observing Right), P1 will play Right in the subgame; hence P2 should play Right herself. Knowing (Right, Right) will be played, P2 will play Across in the first round.

This example shows that observing history *might* plausibly coordinate agents' beliefs on one particular subgame equilibrium. In the pricing game, the same agents act in the morning and afternoon; so it is conceivable that observing pricing decisions in the morning might coordinate beliefs on $y = y^*$ in the afternoon. Appendix A.11 characterizes EFR strategies in this game and shows that this is not so: 'effective demand failures' are rationalizable. Common belief in rationality alone does not stop farmer i expecting $y < y^*$ in the afternoon, and expecting others to share the same pessimistic expectations. True, if i expects $y < y^*$, and acts optimally, she cannot expect others to share her expectations about *prices*, since she wants to set a lower price than the average. But it is perfectly rational to set (e.g.) $p_i = 0.9$, anticipating the average price $p = 1$; since it would

[37] This discussion mirrors Pearce (1984)'s analysis of an essentially equivalent example (game Γ_2 in his paper).

be rational for others to set $p_j = 1$, if they expected $p = 1.1$; and so on. No price profile that i could possibly observe would contradict the hypothesis that all other agents expect $y < y^*$ too.

Learning justifications Another approach to justifying paradoxical equilibria is to view the games as instances of repeated play, and posit a learning process through which agents' beliefs depend on past observation of others' behavior. Consider repeated iterations of a two-stage game where best responses in the 'morning' and 'afternoon' are $x_i = \alpha_{x,x}x + \alpha_{x,y}y$ and $y_i = \alpha_{y,x}x + y$. Agents update beliefs about x and y using decreasing-gain rules $x^e_{t+1} = x^e_t + g_t(x_t - x^e_t)$, $y^e_{t+1} = y^e_t + g_t(y_t - y^e_t)$, where $\{g_t\}_{t=0}^{\infty}$ satisfies $g_t \in (0,1)$, $\lim_{t\to\infty} g_t = 0$, $\sum_{t=1}^{\infty} g_t = \infty$. (Their beliefs about x are only relevant when choosing x_i in the morning; in the afternoon, they observe x.)

First take the case $\alpha_{x,x} = 1$, $\alpha_{x,y} \neq 0$, $\alpha_{y,x} = 0$ (when $\alpha_{x,y} = \kappa > 0$, this nests a repeated version of the static pricing game). Every equilibrium features $y = 0$; as in the pricing game, if $y \neq 0$, agents could not be playing best responses in the morning. But $y = 0$ is not locally stable under learning. If expectations y^e_t initally differ from 0, they can only change if the observed value of y_t changes; but in a pure coordination game where agents set $y_t = y^e_t$, y_t will not change until y^e_t changes. Adjusting expectations x^e_t and behavior x_t to correct past errors, however rapidly, does not bring agents to an equilibrium where they are choosing x_t (e.g., setting prices) optimally.

If instead $\alpha_{x,y}\alpha_{y,x} \neq 0$, the unique equilibrium $(0,0)$ is only locally stable if $\alpha_{y,x}\alpha_{x,y} < \min\{0, 1-\alpha_{x,x}\}$ (see Appendix A.12). y must have indirectly self-stabilizing dynamics that go through x: an increase in y elicits a change in x that reduces y. This may or may not be satisfied in any particular case. Learning does not *generically* rationalize paradoxical equilibria.

Finally, we return to the dynamic pricing game with $T > 0$; since we are interested in asymptotic behavior under learning, we consider its limit as $T \to \infty$. We adopt the 'anticipated utility' approach (Kreps, 1998): farmers update their beliefs about others' behavior using statistical methods, but at any point in time act as if their beliefs will never be revised. In the morning of date t, each farmer forecasts aggregate inflation and consumption at all dates $t + k \geq t$ using the model $\pi_{t+k} = \pi^m_t$, $c_{t+k} = c^m_t$, $\forall k \geq 0$; that is, they expect a constant rate of inflation and level of consumption (expressed in log-deviations from the equilibrium with $\Pi = \beta$, $c = y^*$). In the afternoon of date t, farmers use the same model with beliefs (π^a_t, c^a_t). Farmers update their inflation forecast after observing realized inflation in the morning, but not after observing consumption in the afternoon, and update their consumption forecast after the afternoon, but not after the morning:

$$\pi^a_t = \pi^m_t + g_t(\pi_t - \pi^m_t), \qquad \pi^m_{t+1} = \pi^a_t, \qquad c^a_t = c^m_t, \qquad c^m_{t+1} = c^a_t + g_t(\widehat{c}_t - c^a_t) \tag{13}$$

where π_t and $\widehat{c}_t$ denote realized (log-deviations of) inflation and aggregate consumption, and $\{g_t\}_{t=0}^{\infty}$ is as described above. Appendix A.13 shows that the mapping from beliefs to realized variables is

$$\pi_t = \left[1 + \frac{\beta}{1-\beta}\frac{1}{1+\gamma\varphi}\right]\pi^m_t + \frac{\sigma+\varphi}{1-\sigma+\gamma(\varphi+\sigma)}c^m_t \tag{14}$$

$$\widehat{c}_t = \frac{\beta}{1-\beta}\frac{(1+\gamma\varphi)\sigma^{-1}+\gamma-1}{1+\gamma\varphi+\beta(\gamma-1)\sigma}\pi^a_t + \frac{1+\gamma\varphi-\beta(\gamma-1)\varphi}{1+\gamma\varphi+\beta(\gamma-1)\sigma}c^a_t \tag{15}$$

Appendix A.13 shows that this system is locally unstable. One cannot rationalize equilibrium outcomes in the pricing game by appealing to an evolutive process in which farmers' experiences setting prices and choosing quantities in the past lead them to expect $c = y^*$ in the afternoon subgame. Instead, these experiences would lead them to expect outcomes far from equilibrium.

This should not be surprising, because we are studying dynamics under a nominal interest rate peg (at zero). Expectations of higher inflation can be self-sustaining because they lower the real return on money, encouraging spending today; but since households do not want to work more hours, they raise prices in a mutually self-defeating attempt to set a higher price than their peers, further increasing inflation. This leaves open the possibility that (as in Howitt (1992)) the rational expectations equilibrium would be learnable with interest-bearing money and an 'active' Taylor rule. Note though that (unlike in Howitt (1992)), even with flexible prices, *real quantities* are unstable under 'passive' monetary policy: price flexibility alone does not guarantee full employment.

4.5 What do these examples imply about equilibrium models more broadly?

These examples are special cases, but their key ingredients are standard features of equilibrium models. Similar paradoxes could arise (but would be harder to detect) in more complicated equilibrium models. The conditions required to rule them out are restrictive: finite players and strategies.

But the examples also illustrate a more general point. The assumption that agent i optimizes at date t is not an assumption about i's behavior, or about what happens at date t. In the ij game, the truth of this assumption depends solely on j's behavior at date $t+1$. Claims about i's conduct – arguing that i is smart and wouldn't make systematic errors, or describing how i adjusts his beliefs to correct past errors – cannot support the prediction that i optimizes. It is simply not within i's power to do so: whether his action proves optimal depends entirely on whether j decides, at a later date, to act in ways that rationalize i's decision. To *assume* i optimizes is to assume j acts in this way; to justify that assumption, we need an explanation of *why* j would do so, which an equilibrium model does not provide. In more general models, too, the assumption that i acts optimally asserts that all agents' current and future behavior satisfies a certain dynamic pattern. To justify this assumption, we must explain why *all* agents choose to conform to this pattern, by explicitly modeling the dynamic process through which equilibrium is achieved.

This point is not new. The learning literature has emphasized that learning in self-referential systems differs from single-agent learning because agents are aiming at a moving target, trying to predict each other's nonstationary behavior (Marcet and Sargent, 1992; Young, 2004). Even if each agent i adjusts his own forecasts and behavior to correct past errors, this may not improve aggregate forecast accuracy: i's change in behavior may reduce the accuracy of j's forecasts. Analogously, the older literature on stability of Walrasian equilibrium stressed that even if prices adjust in response to excess demand in each individual market, the whole system might be unstable: a price increase in response to excess demand in market i may disrupt equilibrium in market j (Scarf, 1960; Saari and Simon, 1978). But the point remains under-appreciated; for example, rational expectations is still often viewed as an assumption about *individuals*' cognitive abilities, rather than the relation

between all agents' beliefs and behavior. We return to this theme in Section 6.

5 Paradoxes of immaculate revelation

Rational expectations equilibrium (REE) assumes agents' subjective probability distributions over all variables, conditional on their information sets, equal the objective conditional distributions. When agents' information sets include endogenous variables such as prices, this leads to a second, well-known paradox: prices may reveal information no one is endowed with (Dubey et al., 1987).

The paradox is clearest in the following model. Nature draws $\theta \sim N(0, \sigma^2)$, and agents $i = 1, 2$ each seek to minimize $(a_i - \theta)^2$. Each agent's information set contains *only* the other agent's action; neither agent is directly endowed with any information about the fundamental θ. Suppose for now that a_1, a_2 may be measurable with respect to θ. Then a REE consists of functions a_1, a_2 satisfying

$$a_1(\theta) \in \arg\min_{a_1} \mathbb{E}[(a_1 - \theta)^2 | a_2 = a_2(\theta)], \qquad a_2(\theta) \in \arg\min_{a_2} \mathbb{E}[(a_2 - \theta)^2 | a_1 = a_1(\theta)] \tag{16}$$

In one equilibrium, $a_1(\theta) = a_2(\theta) = 0$ for all θ. Since the other agent's action is uninformative about θ, each agent's action equals its unconditional mean. But in another equilibrium, $a_1(\theta) = a_2(\theta) = \theta$. While agent 1 does not directly observe θ, she can infer it from agent 2's action, and vice versa.[38]

This is absurd. If agent 1 learns θ from agent 2's action, 2 cannot infer θ from a_1; she must already have known θ before 1 acts. Process models satisfying Principle 1 respect this intuition: if the model's graph has an edge from $a_2(\theta)$ to $a_1(\theta)$, it cannot also have one from $a_1(\theta)$ to $a_2(\theta)$ since the graph must be acyclic. But (16) describes an equilibrium model violating Principle 1, which has a cycle, with edges from $a_1(\theta)$ to $a_2(\theta)$ and $a_2(\theta)$ to $a_1(\theta)$. Assumption (16), stating that 1 optimizes given 2's action, does not describe a sequential process, in which 1's action is caused by her information set prior to acting, but posits a simultaneous relation between all agents' beliefs and actions. There is no reason to think *any* possible process would cause this relation to hold.

Following Kreps (1977), this paradox is often ruled out by requiring that endogenous variables are measurable with respect to the information possessed by all agents taken together. In the simple example above, this rules out the 'unrealistic' fully revealing equilibrium. But the Kreps (1977) criterion is easily circumvented, e.g. by introducing a third agent who observes θ and seeks to minimize $(a_3)^2$. Agent 3 always plays $a_3 = 0$ and so, even if agents 1 and 2 observe his action, it cannot reveal 3's private knowledge of θ. Yet according to the Kreps (1977) criterion, since the information possessed by all agents taken together includes θ, we are allowed to consider equilibria in which a_1, a_2, a_3 are measurable with respect to θ, and the fully revealing equilibrium is valid.

The paradox is not easily avoided with further refinements. Suppose again there are only two agents with preferences as above, but now $\theta = \theta_1\theta_2$ where θ_1, θ_2 are i.i.d. $N(0, \sigma^2)$. Agent 1 directly observes θ_1, and 2 directly observes θ_2. The combination of both agent's information perfectly reveals θ; but conditional on any one agent's information, the expectation of θ is zero.

[38] Kreps (1977); Dubey et al. (1987) discuss similar paradoxes in models where agents learn from prices. I first learned this simpler version from Stéphane Dupraz and Sushant Acharya, who each discovered it independently.

The Kreps (1977) criterion entitles us to consider equilibria which are measurable with respect to (θ_1, θ_2), and so $a_1 = a_2 = \theta$ is a valid equilibrium. Yet it is hard to imagine any plausible process in which agents learn θ by observing each others' actions, without explicitly pooling their private information: without also knowing θ_2, knowing θ_1 does not allow agent 1 to improve upon $a_1 = 0$.[39]

While this paradox is not new, the rest of this Section shows it is surprisingly hard to avoid.[40]

5.1 An example in which 'immaculate revelation' is necessary

In the above examples, besides the intuitively unreasonable fully revealing equilibrium, there is a more reasonable equilibrium where endogenous variables are not informative. Thus, we might still hope to find some selection criterion that rules out the unreasonable equilibrium. The following example shows that this may not be possible: sometimes *all* REEs feature 'immaculate revelation'.

There are three agents and two goods; p denotes the price of good 1 and good 2 is the numeraire. Preferences are $u_1(x_1^1, x_1^2; \beta, \varepsilon) = \beta \ln x_1^1 + x_1^2$, $u_2(x_2^1, x_2^2; \beta, \varepsilon) = (3-\beta)(1+\varepsilon)\ln x_2^1 + x_2^2$ and $u_3(x_3^1, x_3^2; \beta, \varepsilon) = x_3^2$. x_i^j denotes i's consumption of good j. Agent i has a constant endowment ω_i^j of good j, where $\sum_{i=1}^3 \omega_i^1 = 3$. $\beta \in \{1, 2\}$ with equal probability; $\varepsilon \in \{-\overline{\varepsilon}, \overline{\varepsilon}\}$ with equal probability and is independent of β, where $\overline{\varepsilon} \in (0, 1/2)$. Agent 1 is perfectly informed about β, agent 2 is uninformed, and agent 3 is perfectly informed about ε. Letting $\sigma_i(\beta, \varepsilon)$ denote i's information, a REE is a collection $\{x_i^j(\beta, \varepsilon)\}_{i=1,2,3, j=1,2}$ and a price scheme $p(\beta, \varepsilon)$ such that, for all β, ε:

1. for each agent $i = 1, 2, 3$, $(x_i^1(\beta, \varepsilon), x_i^2(\beta, \varepsilon))$ maximizes $\mathbb{E}\left[u_i(x_i^1, x_i^2; \beta, \varepsilon) \mid p = p(\beta, \varepsilon), \sigma_i(\beta, \varepsilon)\right]$ subject to $p(\beta, \varepsilon)(x_i^1 - \omega_i^1) + x_i^2 - \omega_i^2 \leq 0$,
2. each market $j = 1, 2$ clears: $\sum_i (x_i^j(\beta, \varepsilon) - \omega_i^j) = 0$

We note first that agent 3 does not trade in any equilibrium: $x_3^1 = 0$, $x_3^2 = \omega_3^2$.[41]

In any REE, ε must be revealed. Suppose by contradiction that there is an equilibrium in which ε is not revealed, i.e. $p(\beta, \varepsilon)$ does not depend on ε. Then either β is revealed or not.[42] If it is not revealed (i.e. $p(\beta, \varepsilon)$ is constant), 1's demand for good 1, $x_1^1(\beta, \varepsilon) = \beta/p$, depends on β, but 2's demand does not (since he does not observe it). So the same price cannot clear markets in all states, a contradiction. If β is revealed, demands are $x_1^1 = \beta/p$, $x_1^2 = (3-\beta)/p$, $x_3^1 = 0$, and so the market clearing price is $p(\beta, \varepsilon) = 1$, a constant, which contradicts β being revealed.

To see that there is an equilibrium in which ε is revealed, note that in this case market clearing becomes $\beta/p + (3-\beta)(1+\varepsilon)/p = 3$, so the equilibrium price is $p = 1 + \frac{\varepsilon(3-\beta)}{3} \in \{1 - 2\overline{\varepsilon}, 1 - \overline{\varepsilon}, 1 + \overline{\varepsilon}, 1 + 2\overline{\varepsilon}\}$, which is different for each realization of β, ε: the equilibrium is indeed fully revealing.

[39] Jordan (1982) describes a process with multiple rounds of trading (first suggested by Reiter) which may or may not converge to REE in exchange economies. In the first round traders only condition their excess demands on private information; this information may be revealed in prices, leading to a second round of trading where traders condition their excess demands on private information and information revealed in the first round; and so on. The limit of this process need not be a REE, since the final price function may not include information revealed by earlier prices.

[40] This particular paradox can be avoided by writing explicit sequential games satisfying Principles 3 and 4, rather than market clearing models (Dubey et al., 1987). Such games can still exhibit the paradox discussed in Section 4.

[41] He is only introduced to allow equilibria where p is measurable with respect to ε without violating Kreps (1977).

[42] In this case the economy reduces to Mas-Colell et al. (1995)'s (Example 19.H.3, p722) version of Kreps (1977)'s example, and their nonexistence proof shows that such an REE cannot exist. I repeat the proof here for convenience.

5.2 Partial revelation in noisy rational expectations equilibrium

A common approach to study learning from prices while avoiding some REE paradoxes has been to introduce noise, so that prices partially, but not fully, reveal informed traders' information (Grossman and Stiglitz, 1980). But this does not avoid the paradox that prices can reveal information that no trader possesses, as the following example (which draws on Pálvölgyi et al. (2017)) shows.[43]

The asset market consists of noise traders and rational but uninformed traders. The supply of the risky asset is constant and equals y, and it pays off $\theta \sim N(\overline{\theta}, \sigma^2)$ units of the consumption good; the safe asset has unlimited supply and pays off 1 unit for sure. There is a signal $s = \theta + \varepsilon$ where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, but *no* agents observe it.[44] Noise trader demand for the asset is $u \sim N(0, \sigma_u^2)$. θ, ε and u are independent. Rational traders choose their position x to maximize the expected value of $-\frac{1}{\alpha}\exp(-\alpha x(\theta - p))$ conditional on their information set, which consists of the realization of the price p. That is, a rational expectations equilibrium is a pair $x(s,u), p(s,u)$ such that for all (s,u), $x(s,u)$ maximizes $\mathbb{E}[-\frac{1}{\alpha}\exp(-\alpha x(\theta - p))|p = p(s,u)]$ and $x(s,u) + u = y$.

There exists a non-revealing equilibrium, in which prices depend on u but not s. Traders thus maximize the unconditional expectation $\mathbb{E}[-\frac{1}{\alpha}\exp(-\alpha x(\theta - p))]$, yielding the optimality condition $\overline{\theta} - p - \frac{\alpha}{2}\sigma^2 x = 0$. Market clearing implies $x + u = y$, so the uninformed price is $p^{NRE}(s,u) = \overline{\theta} - \frac{\alpha}{2}\sigma^2(y - u)$. The magenta dashed line in Figure 5 shows $p^{NRE}(s,u)$ as a function of u in a numerical example. To the extent that the asset is risky and traders are risk averse, p is lower when noise trader demand is weak (u is low) and rational traders must absorb more of the supply.

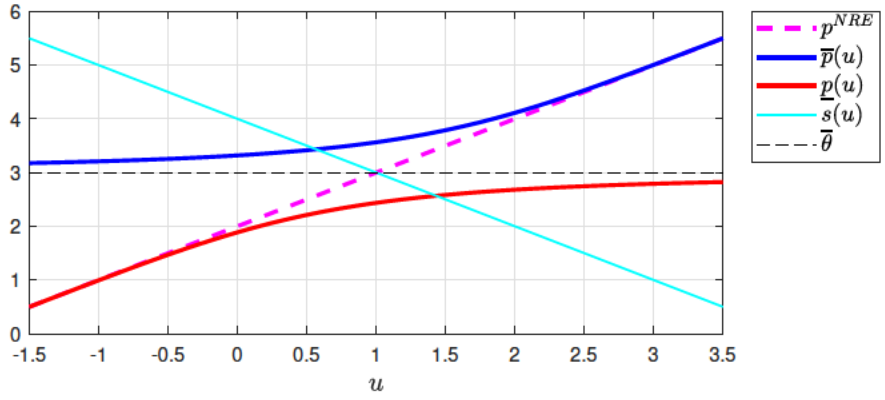


Figure 5: Non-revealing and partially revealing equilibria ($\sigma = \sigma_\varepsilon = \alpha = y = 1, \overline{\theta} = 3$)

However, there also exists a partially revealing equilibrium: Given u, there are two possible prices, $\overline{p}(u)$ (blue line in Figure 5) and $\underline{p}(u)$ (red line), where $p = \overline{p}(u)$ iff $s > s(u)$ and $s(u)$ (cyan line) is a cutoff. The images of $\overline{p}(\cdot)$ and $\underline{p}(\cdot)$ do not overlap. Thus, upon observing p, the trader can infer both u, and whether s is greater or less than the cutoff $s(u)$: the price partially reveals s,

[43] Pálvölgyi et al. (2017) construct discontinuous equilibria in a noisy REE model with some informed traders, in which, besides the information revealed in the standard linear equilibrium, prices also reveal which 'regime' s belongs to. My example, and proof in Appendix A.14, follow their analysis, but consider the limit with *no* informed traders.

[44] Again, to formally satisfy Kreps (1977), we could introduce a third set of agents who observe s but do not trade.

even though no trader observes this signal. Specifically, the equilibrium price function is

$$p(s,u) = \begin{cases} \underline{p}(u) := \overline{\theta} - \alpha\sigma^2(y-u) - \frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi(\alpha\sigma_s(y-u))}{\Phi(\alpha\sigma_s(y-u))} & \text{if } s \leq \overline{\theta} + \alpha\sigma_\varepsilon^2(y-u) := s(u) \\ \overline{p}(u) := \overline{\theta} - \alpha\sigma^2(y-u) + \frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi(\alpha\sigma_s(y-u))}{1-\Phi(\alpha\sigma_s(y-u))} & \text{if } s > s(u) \end{cases}, \sigma_s^2 := \sigma^2 + \sigma_\varepsilon^2$$

If and how prices coordinate economic activity, and aggregate and transmit dispersed information, are core questions in economics (Hayek, 1945). But while REE models show how prices *transmit* information, they do not specify how prices are determined, violating Principles 3 and 4 (no agent explicitly chooses prices, given their prior information). Thus they cannot explain how information gets into prices in the first place (Dubey et al., 1987). Equilibrium analysis *assumes* coordination – agents all agree on what they would do, given any realization of fundamentals – rather than describing the process by which coordination is achieved. Thus, if markets *do* aggregate information and coordinate activity, equilibrium models cannot explain how; and sometimes, as in the paradoxes above, these models predict that markets aggregate and coordinate when they could not possibly do so. Instead, studying these questions requires writing explict process models.

6 Equilibrium 'alternatives' to rational expectations

The rational expectations (RE) assumption that agents' subjective probability distributions coincide with 'objective' probability distributions is often criticized, both because it seems to credit people with unrealistic cognitive abilities or understanding of their environment, and because it conflicts with empirical evidence, e.g. regarding the predictability of forecast errors. This has led economists to explore various alternatives to RE which seem to provide a more realistic description of belief formation. Many such alternatives assume that there is still *some* relation between subjective and objective probability distributions, but the two need not coincide (Woodford, 2013).

But RE is not a description (realistic or otherwise) of the way in which individuals form beliefs. It is an equilibrium condition: an assumption about the *relation* between an individual's beliefs and others' behavior, posited without describing a process which causes this relation to hold. Assuming i has rational expectations about j's behavior constrains j's behavior as much as it constrains i's beliefs. Even if each individual perfectly understood her environment and had superhuman cognitive abilities, rational introspection might not lead them all to have mutually consistent expectations. Conversely, many alternatives to RE still assume a relation between i's belief and j's behavior – albeit a 'distorted' one – without describing a process causing this relation to hold. Thus, as I now show, they are still equilibrium assumptions, and exhibit the same paradoxes as RE models.

Several popular departures from full information RE can be classified as follows. One approach maintains RE, but assumes agents only observe noisy signals of exogenous or endogenous variables (which may be chosen optimally subject to constraints, as in the rational inattention (Sims, 2003) literature); this is still an equilibrium assumption. In the second approach (e.g. diagnostic expectations (Bordalo et al., 2018), cognitive discounting (Gabaix, 2020)), agents' subjective distributions are assumed to be related to, but not identical to, the objective distribution of the same variables:

also an equilibrium assumption. In the third approach, agents observe exogenous variables and reason correctly given incorrect beliefs about how *other* agents reason and behave. This may or may not be an equilibrium assumption depending on how these incorrect beliefs relate to actual behavior: while 'shallow reasoning' (Angeletos and Sastry, 2021) retains equilibrium, level-k (Crawford et al., 2013) is a non-equilibrium framework, closer to the eductive viewpoint discussed in Section 4.4. I discuss illustrative examples of each approach; the points apply more broadly.

6.1 Rational inattention

As an example of the first approach, we introduce rational inattention into a simple game of the kind discussed in Section 4.3. A separate unit continuum of agents chooses actions at dates 0 and 1; for tractability, these actions are binary, $\in \{0,1\}$. An exogenous shock z, equals z_L or z_H with equal probability, $0 < z_L < z_H < 1$. Date 0 agents ('attackers') receive payoff $v(y)x_i$ from choosing $x_i \in \{0,1\}$, where y is the fraction of date 1 agents ('defenders') choosing $y_i = 1$ ('defend') and $v(y) = \ln((1-y)/y)$. Defenders receive payoff $y_i(p(z) - z)$ from choosing $y_i \in \{0,1\}$, where $p(z)$ is the fraction of date 0 agents who choose $x_i = 1$ when the state is z. Attackers want to attack ($x_i = 1$) when defenders do *not* defend. Defenders want to defend when attackers *do* attack; it is only worth doing so when the fraction of attackers is weakly greater than the cost of defending, z.

Attackers do not observe z directly, but can pay a cost to acquire a noisy signal s which is correlated with z with joint distribution $F(s,z)$. Following the rational inattention literature, we assume the entropy-based cost function $C(F,G) = \lambda(H(G) - \mathbb{E}_s[H(F(\cdot|s))])$ where $\lambda \geq 0$ is the unit cost of information, G the prior and $H(P) = -\sum_k P_k \ln P_k$ the entropy of a discrete probability distribution. Equivalently, attackers choose state-contingent choice probabilities subject to a cost based on the mutual information between states and actions (Matějka and McKay, 2015):

$$\max_{(p(z_L),p(z_H))\in[0,1]^2} \frac{1}{2}p(z_L)v(y(z_L)) + \frac{1}{2}p(z_H)v(y(z_H)) - C(p,G)$$

This is an equilibrium condition: the date 0 variables $p(z_L), p(z_H)$ depend on the the proportion of defenders defending in each state at date 1 $y(z_L), y(z_H)$, violating the arrow of time. Appendix A.15 solves for $p(z_j)$. With free information ($\lambda = 0$), attackers all attack ($p(z_j) = 1$) when the net benefit $v(y(z_j))$ is positive, and all defend ($p(z_j) = 0$) when it is negative; when $v(y(z_j)) = 0$, they are indifferent, and any $p(z_j) \in [0,1]$ is optimal. As the information cost λ increases, they acquire noisier signals and make more mistakes, pushing $p(z_L)$ and $p(z_H)$ closer together for a given $y(\cdot)$.

Defenders observe the fraction of attackers attacking $p(z_j)$ before deciding whether to defend. If $p(z_j) > z_j$ in state j, they all defend ($y(z_j) = 1$); if $p(z_j) < z_j$, none of them defend ($y(z_j) = 0$); if $p(z_j) = z_j$, they are indifferent, and any $y(z_j) \in [0,1]$ is consistent with optimal behavior.

In equilibrium, it must be that defenders are indifferent and $y(z_j) \in (0,1)$. If by contradiction e.g. defenders always defend in state j ($p(z_j) \geq z_j$, $y(z_j) = 1$), attackers get $-\infty$ if they attack in

this state, so they would willingly pay higher costs to avoid doing so, reducing $p(z_j)$. We have

$$y(z_j) = \frac{\left(\frac{\overline{z}}{1-\overline{z}}\right)^{\lambda}}{\left(\frac{z_j}{1-z_j}\right)^{\lambda} + \left(\frac{\overline{z}}{1-\overline{z}}\right)^{\lambda}},\ j \in \{L, H\} \text{ where } \overline{z} = \frac{z_L + z_H}{2}.$$

If information is free ($\lambda = 0$), the equilibrium has mixed strategies. While attackers are always indifferent, in equilibrium they must attack more often when defense is more costly ($p(z_H) = z_H > z_L = p(z_L)$) to keep defenders indifferent at date 1. While defenders are always indifferent, in equilibrium they must randomize with equal probability $y(z_L) = y(z_H) = 1/2$; otherwise, attackers would not have been indifferent at date 0. This is paradoxical: at date 1, attackers have already moved, and defenders are indifferent. In a one-shot game, what ensures they play $y = 1/2$?

Rational inattention ($\lambda > 0$) does not remove the paradox. While it remains true that defenders are always indifferent, now, in equilibrium, they must defend with higher probability when z is low than when it is high. This ensures that attackers acquire some information about the state, and attack more frequently when z is high and there are fewer defenders (which is necessary to keep defenders indifferent and willing to randomize). But again, at date 1, attackers have already moved, and defenders are indifferent. What ensures they randomize in precisely those proportions that would have incentivized 'correct' play by attackers, had they been anticipated at date 0?

Under full information, the assumption that attackers' behavior is optimal at date 0 pins down defenders' behavior at date 1. Under rational inattention, the assumption that attackers' behavior is *constrained*-optimal does the same thing. The paradox comes from *assuming* a relation between agents' behavior at different dates, even a 'distorted' one.

6.2 Diagnostic expectations and cognitive discounting

Departures from RE which directly posit a 'distorted' relation between subjective and objective distributions – diagnostic expectations (Bordalo et al., 2018) and cognitive discounting (Gabaix, 2020) – also suffer from the paradoxes in Section 4. Since these approaches change how agents forecast the future, it is convenient to consider an infinite-horizon rather than 2-period example.

At date $t = 0, 1, ...$, a continuum of firms choose how many vacancies v_t to post to maximize $v_t(z_t + \beta \widetilde{E}_t J_{t+1})$, where $J_{t+1} = z_{t+1} - \eta v_{t+1}$ denotes the continuation value of a filled vacancy, and $\widetilde{E}_t J_{t+1}$ denotes firms' (potentially non-rational) expectation of this variable. z_t follows the AR(1) process $z_t = \rho z_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, \sigma^2)$, $\rho \in (0, 1)$. Interpreting z_t as match productivity, the parameter $\eta > 0$ crudely represents the idea that a tighter labor market in the future reduces the value of a filled vacancy, either by increasing workers' outside options and the wage they can extract, or by increasing the probability that matched workers find another job and leave the firm.

In any equilibrium, the objective relation between J_{t+1} and z_t can be written $J_{t+1} = \mu(z^t) + \delta\varepsilon_{t+1}$, where $\mu(\cdot)$ and δ are endogenous. Under **diagnostic expectations** (DE), firms' subjective probability distribution over J_{t+1}, given history $\widehat{z}^t$, does not coincide with the objective distribution $h(\widehat{J}_{t+1}|\mu(z^t) = \mu(\widehat{z}^t))$. (I follow Bianchi et al. (2024)'s implementation of DE in general equilib-

rium; hats denote realizations of random variables.) Instead, firms overestimate the probability of outcomes which are more likely to occur given the realized state $z_t = \widehat{z}_t$ than under the 'reference event' in which there are no shocks, $z_t = \mathbb{E}_{t-1} z_t = \rho \widehat{z}_{t-1}$. This yields the diagnostic distribution

$$h_t^\theta(\widehat{J}_{t+1}) := h(\widehat{J}_{t+1}|\mu(z_t) = \mu(\widehat{z}_t) \left(\frac{h(\widehat{J}_{t+1}|\mu(z_t) = \mu(\widehat{z}_t))}{h(\widehat{J}_{t+1}|\mu(z_t) = \mathbb{E}_{t-1}\mu(z^t))} \right)^\theta \frac{1}{a}$$

where a is a normalizing constant, and the diagnosticity parameter $\theta > 0$ measures the degree to which beliefs are distorted by the representativeness heuristic. Since the objective distribution of J_{t+1} is normal, the diagnostic distribution is also normal with mean $\mathbb{E}_t^\theta[J_{t+1}] = \mu(z^t) + \theta[\mu(z^t) - \mathbb{E}_{t-1}\mu(z^t)]$. Assuming DE ($\widetilde{E}_t J_{t+1} = \mathbb{E}^\theta J_{t+1}$), the firm's problem yields the optimality condition $z_t + \beta \mathbb{E}_t^\theta J_{t+1} = 0$. This implies J_{t+1}'s objective mean is $\mathbb{E}_t J_{t+1} = -\frac{1}{\beta(1+\theta)} z_t - \frac{\rho\theta}{\beta(1+\theta)} z_{t-1}$, i.e.

$$\mathbb{E}_t v_{t+1} = \frac{1}{\eta} \left(\rho + \frac{1}{\beta(1+\theta)} \right) z_t + \frac{\rho\theta}{\beta(1+\theta)} z_{t-1}$$

First take the case of RE, $\theta = 0$. In any equilibrium, at dates t and $t+1$, firms must be indifferent between posting any number of vacancies. Suppose a positive productivity shock $z_t > 0$ realizes at date t. This tends to increase the return to vacancy posting; to keep date t firms indifferent and prevent them from posting infinite vacancies, it must be that v_{t+1} increases (i.e. date $t+1$ firms post more vacancies), reducing the continuation value of a filled vacancy J_{t+1}. But at date $t+1$, firms are indifferent. Why should they conform to this equilibrium and post more vacancies just because *last period's* productivity was high (even if $\rho = 0$, and this has no effect on current productivity)? Again, the assumption that date t firms optimize given rational expectations about date $t+1$ firms' behavior constrains the behavior of not only date t firms, but also date $t+1$ firms.

Now assume DE, $\theta > 0$. Again, in equilibrium the value of a filled vacancy J_{t+1} must fall when z_t rises to keep date t firms indifferent. More specifically, firms' *diagnostic expectation* $\mathbb{E}_t^\theta J_{t+1}$ must be decreasing in z_t, and unrelated to $z_{t-1}, z_{t-2}, ...$ But diagnosticity naturally makes $\mathbb{E}_t^\theta J_{t+1}$ depend on z_{t-1}: if z_t is below its expected value ρz_{t-1}, firms over-extrapolate this shock into the future, underpredicting z_{t+1}. Thus, if J_{t+1} were *objectively* independent of z_{t-1}, a higher z_{t-1} would increase $\mathbb{E}_t^\theta z_{t+1}$, and lower $\mathbb{E}^\theta J_{t+1}$, which cannot be the case in equilibrium since $\mathbb{E}^\theta J_{t+1}$ can only depend on z_t. The only way to offset the dependence of $\mathbb{E}_t^\theta J_{t+1}$ on z_{t-1} is for J_{t+1} to be objectively *increasing* in z_{t-1}. Then for a given value of z_t, following higher z_{t-1}, the 'kernel of truth' in firms' forecasts leads them to predict lower J_{t+1}, over-extrapolation leads them to predict higher J_{t+1}, and on net their forecast is unchanged. Arguably, DE *worsens* the paradox: Even though date $t+1$ firms are indifferent between any level of vacancy posting v_{t+1}, in equilibrium v_{t+1} must depend on not only z_t, but z_{t-1} – even though z_{t-1} is irrelevant for predicting all variables from $t+1$ onwards.

The same paradox arises with **cognitive discounting** (CD) (Gabaix, 2020). Now assume that while z_t objectively follows the AR(1) process described above, firms perceive that $z_{t+1} = m(\rho z_t + \varepsilon_{t+1})$, where $m \in [0, 1]$. Firms correctly understand the (endogenous) mapping from z^t to J_t; but when $m < 1$, inattention to the future leads them to underestimate fluctuations in z_{t+1}.

Their optimality condition becomes $z_t = \beta m \mathbb{E}_t J_{t+1}$, implying $\mathbb{E}_t v_{t+1} = \eta^{-1}(\rho + (\beta m)^{-1}) z_t$. Relative to RE ($m = 1$), with CD ($m < 1$), date $t+1$ vacancies are more sensitive to date t productivity. When z_t increases, in equilibrium date t firms' expected continuation value must fall to prevent them from posting infinite vacancies. But since firms underestimate changes in future variables, the *true* continuation value must fall even more, relative to RE, to induce the required change in expectations. This requires date $t+1$ vacancies to increase even more than in RE. Again, since date $t+1$ firms are indifferent, it is intuitively unclear why they would respond to z_t at all.

DE and CD also permit the 'immaculate revelation' paradoxes discussed in Section 5. Suppose overlapping generations of investors live for two periods, consume when old, and allocate their endowment between a safe asset in unlimited supply yielding 1 unit of consumption, and risky capital k_t, which has price q_t, pays z_{t+1} and fully depreciates. A date t investor chooses k_t to maximize $E_t^i\left[-\frac{1}{\alpha}\exp(-\alpha(z_{t+1} - q_t)k_t)\right]$, where E_t^i denotes their potentially non-rational expectation at t. Date t investors observe current and past prices $q^t = q_t, q_{t-1}, ...$ No one observes shocks z^t. Assuming that after observing q^t, the investor perceives z_{t+1} to be normally distributed with mean $E_t^i[z_{t+1}]$ and variance $V_t^i[z_{t+1}]$, their demand for capital satisfies $k_t = (\alpha V_t^i[z_{t+1}])^{-1}(E_t^i[z_{t+1}] - q_t)$ (see Appendix A.16). Capital is produced by firms who choose k_t to maximize $q_t k_t - \frac{\nu}{2}k_t^2$, yielding the supply curve $k_t = \frac{q_t}{\nu}$. Thus the equilibrium price is $q_t = (\nu E_t^i[z_{t+1}])/(\alpha V_t^i[z_{t+1}] + \nu)$.

Under **RE**, there is a fully revealing equilibrium with $E_t^i[z_{t+1}] = \rho z_t$, $V_t^i[z_{t+1}] = \sigma^2$, and $q_t = \rho z_t/(\alpha\sigma^2 + \nu)$. Given this price function, investors can infer z_t after observing q_t; given knowledge of z_t and a rational forecast of z_{t+1}, their demands are such that q_t is indeed an equilibrium.

Under **CD**, investors perceive that the state evolves according to $z_{t+1} = m(\rho z_t + \varepsilon_{t+1})$. There is still a fully revealing equilibrium, in which $q_t = m\rho z_t/(\alpha m^2\sigma^2 + \nu)$. Given this price function, investors can infer z_t from q_t. The *way* in which they use this information to forecast z_{t+1} exhibits systematic errors. But because this forecast retains *some* relation to the rational one, and hence z_t, investors' demand for capital moves with z_t, and the market-clearing price remains fully revealing.

Under **DE**, investors perceive that the state evolves according to $z_{t+1} = (1+\theta)\rho z_t - \theta\rho^2 z_{t-1} + \varepsilon_{t+1}$. Thus, in the fully revealing equilibrium, the price is $q_t = (\rho(1+\theta)z_t - \theta\rho^2 z_{t-1})/(\alpha\sigma^2 + \nu)$. Given this equilibrium price function, investors can infer z_t from the whole past history of prices:

$$z_t = \frac{\alpha\sigma^2 + \nu}{\rho(1+\theta)} \sum_{\ell=0}^{\infty} \left(\frac{\theta\rho}{1+\theta}\right)^{\ell} q_{t-\ell} \tag{17}$$

Investors use the information in q^t to forecast z_{t+1}, and their demand for capital reflects this forecast. Although the forecast is systematically incorrect, it nonetheless comoves with z_t, and this embeds enough information into the market-clearing price that the latter is fully revealing.

6.3 'Shallow reasoning' and level-k beliefs

We return to the IOR game and depart from RE by assuming '**shallow reasoning**' (Angeletos and Sastry, 2021; Angeletos and Lian, 2023).[45] Each trader knows θ, and chooses C_i, B_i optimally

[45] Throughout this section, for simplicity we consider the limiting case of this game where $\delta \to 0$.

given her expectations about others' behavior, but believes only a fraction λ of others know θ, while the remaining $1-\lambda$ are uninformed and act as in the $\theta = 1$ equilibrium. Each trader correctly anticipates how other informed traders behave, but fails to realize *all* traders are in fact informed.

The characterization of individually optimal behavior in Section 4.2 remains valid. In the afternoon subgame, traders bid $B_i = \overline{B}$ if $p < 1$, $B_i = 0$ if $p > 1$, and are indifferent if $p = 1$. In the morning subgame, their bid C_i satisfies (11) given their expectations p, B about the aggregate price p and bond market bid B. Now, however, these expectations are given by $p = \lambda C(\theta) + (1-\lambda)C(1)$, $B = \lambda B(\theta) + (1-\lambda)B(1)$, where $C(\theta), B(\theta)$ denote i's (correct) beliefs about the bids of another informed trader, and $C(1) = B(1) = 1$ denote i's beliefs about the bids of an uninformed trader.

Shallow reasoning does not resolve the paradox; if anything, it intensifies it. Traders' bids in the afternoon subgame are now even more responsive to the value of θ in the morning subgame: $B_i = \lambda^{-1}(\theta^{-1} - 1) + 1$.[46] In equilibrium, traders must expect $1 + r = \theta$, i.e. they must expect the bids placed in the bond market to be lower when θ is high, and higher when θ is low. Since they incorrectly believe a fraction $1-\lambda$ of traders do not adjust their bid at all, they must believe that those who *do* adjust, adjust by more, to compensate for inattentive traders' inertia. And since by assumption they correctly anticipate how informed traders behave, informed traders must in fact be highly responsive to θ. But again, any profile of B_i is an equilibrium of the afternoon subgame.

The RE assumption that trader i optimizes given correct beliefs about others' play does not just constrain i's behavior, but also others' future behavior. Assuming that i optimizes given beliefs that have some *distorted* relation to the way other agents actually behave still restricts how other agents behave. Shallow reasoning still implies an equilibrium rather than a process model: the model's graph has cycles and does not respect the arrow of time ($C_i \leftarrow B$), violating Principle 1.

All these deviations from REE make agents' expectations about exogenous or endogenous variables less sensitive to changes in the true values of these variables. One might expect this to mitigate paradoxes – seemingly driven by unrealistic 'GE effects' – in which agents' behavior responds to shocks which, intuitively, should not affect their decisions. This prediction fails for the same reason that the direct-indirect effect decompositions in Section 3 suggest misleading predictions about the effect of 'damping' particular channels. If an agent becomes less attentive to aggregate variables, in equilibrium these variables may simply respond more to shocks to induce her to respond as 'required'. Her optimality condition constrains others' behavior as much as her own; 'damping' this optimality condition alters others' behavior, and can make it *more* responsive to shocks.

Contrast this to a superficially similar approach which *does* depart from equilibrium and does avoid the paradoxes: **level-k reasoning**. Level-0 ($L0$) agents take some 'baseline' action (C^0, B^0). An $L1$ agent chooses C^1 optimally given θ and the initial assumption that all other agents are $L0$ and will play (C^0, B^0). Given our extensive form game, her choice of B will not depend on these prior beliefs, but only on the realized value of C; let B_1 be her optimal choice if all other agents are $L1$ and play C^1. An $L2$ agent initially assumes all other agents are $L1$ and will play (C^1, B^1)... That is, C^k, B^k are defined recursively: $C^{k+1} = \frac{\theta}{1+\theta}\left(C^k + B^k + \left(\frac{1-C^k}{C^k}\overline{B}\right)^+\right)$, and

[46] See Appendix A.17 for all proofs relating to the shallow reasoning and level-k examples in this section.

$B^{k+1} = \overline{B}$ if $C^{k+1} < 1$, $= 0$ if $C^{k+1} > 1$, and can take any value if $C^{k+1} = 1$. The distribution of 'cognitive depth' k in the population is $\{\lambda_k\}_{k=0}^{\infty}$, $\sum_{k=0}^{\infty} \lambda_k = 1$. Aggregate actions are given by $C = \sum_{k=0}^{\infty} \lambda_k C^k$, $B = \sum_{k=0}^{\infty} \lambda_k \widetilde{B}^k$ where $\widetilde{B}^0 = B^0$ and $\widetilde{B}^k$ is a best response to C for all $k > 0$.

The game theoretic literature on level-k beliefs (Crawford et al., 2013) typically assumes $L0$ behavior to be nonstrategic (e.g. a uniform random distribution over all strategies). This produces a process model satisfying the principles in Section 2: Lks' behavior is determined by their beliefs about $L(k-1)$s' behavior, which in turn are recursively determined by their beliefs about $L0$s, but no relation, not even a distorted one, is assumed between these beliefs and actual behavior (the model's graph is acyclic). As such, the paradox does not arise. Whereas in Nash equilibrium, $C = 1$ and $B = \theta^{-1}$, even though agents are indifferent between choosing any value of B in the afternoon market, with level-k, $C \neq 1$ and B_k equals either 0 or $\overline{B}$ for all $k > 0$, implying $B \neq \theta^{-1}$.[47]

The difference is that, like the eductive and learning approaches discussed in Section 4, level-k specifies a *recursive* process through which agents form beliefs about others' behavior. If one wants to avoid paradoxes and other undesirable features of REE, these approaches are more promising than assuming a simultaneous, albeit 'distorted', relation between beliefs and behavior.

7 Conclusion

The inconsistency between equilibrium and causal process creates problems. Even if we trust equilibrium models, attempts to understand them by decomposing their mechanisms suggests misleading predictions about how the models work. And in some cases, equilibrium assumptions generate predictions so contrary to *any* plausible causal process – history determines the future without affecting any intervening state variables; information no one knows can get into prices – that we should distrust the assumptions. The possibility of such paradoxes indicates the need to check whether models with explicit descriptions of plausible causal processes support equilibrium assumptions more generally. Such models are also worth studying in their own right.

This paper's focus has been critical rather than constructive: I have not constructed many economic examples of process models. But there is no technical barrier to doing so. One route is to take a market game and replace Nash equilibrium with a model of belief formation. Of course, this retains the assumption of optimizing behavior, despite abandoning equilibrium: one may think it preferable to try and specify behavioral rules which accurately describe household and firm decision-making, informed by qualitative and quantitative data, without presuming optimization. This is an important question, but logically separate from that of equilibrium versus process.

Attempts to model disequilibrium face familiar criticisms. Doing so introduces free parameters: there are many models of non-equilibrium learning, and economic theory does not single out any one of them. This problem is not insuperable: learning models can be disciplined using empirical

[47] Macroeconomic applications of level-k (Farhi and Werning, 2019; Iovino and Sergeyev, 2023) typically assume $L0$ agents behave as in the baseline REE without shocks ($\theta = 1$), $C_0 = B_0 = 1$. Arguably then, this still produces an equilibrium model, which does not specify the process which leads $L0$ agents to play $B_0 = 1$ even though other actions would be optimal if $C = 1$. But, at least in this example, the response of B to θ remains markedly different from the Nash equilibrium. When $\theta \neq 1$, again, $C \neq 1$, $B_k \in \{0, \overline{B}\}$ for all $k > 0$, and B has no relation to θ^{-1}.

evidence on belief formation (e.g. survey data on expectations), just as the specification of tastes and technology in equilibrium models is informed by data. The more serious criticism is that belief formation may be much less stable than tastes and technology, particularly in response to policy interventions. Even if a 'learning rule' accurately describes belief formation in some sample, assuming people continue to mechanically follow that rule following an intervention could yield incorrect predictions (Lucas, 1976); we can judgmentally modify the rule to avoid 'implausible' predictions, but that introduces some arbitrariness. But while assuming equilibrium reduces this arbitrariness, there is no reason to think it will produce more accurate predictions. Just because an equilibrium model correctly observes *that* an intervention changes behavior does not mean the model correctly describes *how* it does so. Why should we expect REE to describe behavior unless a range of plausible learning processes converge to REE? Even if they *do* eventually converge, given a stationary environment, how does that justify the widespread use of REE to model *short-run* dynamics? If equilibrium models as a class had a stellar record of predicting the response to interventions, we might set these concerns aside. The available evidence suggests they do not.[48]

My case for *theoretical* process models is not an argument against using simultaneous equations in applied work. Even if reality is well-described by a recursive process model, we rarely observe all relevant variables at a high enough frequency to estimate it directly, and time aggregation introduces simultaneity (Bentzel and Hansen, 1954). In such cases, one can still write a theoretical process model and account for time-aggregation when taking the model to data. Directly assuming equilibrium will not provide a good approximation to this time-aggregated model unless the true process converges rapidly to equilibrium under all possible interventions. This is not guaranteed, and can only be checked by explicitly modeling this process (Fisher, 1970; Bongers et al., 2022).

Finally, explicitly modeling how prices are determined, beliefs are formed, etc. is more complicated than just assuming equilibrium. Surely our comparative advantage as economists is using a few general principles to *avoid* modeling all this detail? Process models offer some compensating advantages: their recursive structure means they can be solved forward by simple iteration without fixed-point calculations, and (cf. Section 3) makes them easier to understand causally. But it is undeniable that modeling disequilibrium process will take additional work, even if, in many cases, we ultimately find that the process does converge to equilibrium. Ironically, the reward for this labor is similar to that promised by the partisans of the microfoundations revolution: a deeper, more robust understanding of how aggregate relations (heretofore simply *assumed*) arise from intentional human action. Paraphrasing Lucas (1980): if we can describe the causal process which causes the assumptions of market clearing and consistent beliefs to hold, we will know what these assumptions mean, we will understand them in a sense that equilibrium models will *never* be understood. This is exactly why we care about the disequilibrium foundations of equilibrium economics.

[48] Applied general equilibrium models frequently make incorrect predictions about the effect of liberalization on the level and pattern of trade (Kehoe, 2005; Kehoe et al., 2017). Merger simulations often make inaccurate predictions of the effect of mergers on pricing behavior (Peters, 2006). Outside of the experimental economics literature, such attempts to validate equilibrium models' predictions about the effect of policy interventions are surprisingly rare. "A search of the structural estimation literature reveals few attempts to validate structural models using quasi-experiments" (Keane, 2010), and most of these are optimizing models of individual behavior, not equilibrium models.

References

Adlam, Emily, "Spooky action at a temporal distance," *Entropy*, 2018, *20* (1), 41.

Albert, Jeffrey M. and Suchitra Nelson, "Generalized causal mediation analysis," *Biometrics*, 2011, *67* (3), 1028–1038.

Angeletos, George-Marios and Chen Lian, "Dampening general equilibrium: incomplete information and bounded rationality," in Rüdiger Bachmann, Giorgio Topa, and Wilbert van der Klaauw, eds., *Handbook of Economic Expectations*, Elsevier, 2023, pp. 613–645.

_ **and Karthik A. Sastry**, "Managing expectations: Instruments versus targets," *The Quarterly Journal of Economics*, 2021, *136* (4), 2467–2532.

Auclert, Adrien, Bence Bardóczy, Matthew Rognlie, and Ludwig Straub, "Using the sequence-space Jacobian to solve and estimate heterogeneous-agent models," *Econometrica*, 2021, *89* (5), 2375–2408.

Bassetto, Marco, "A game–theoretic view of the fiscal theory of the price level," *Econometrica*, 2002, *70* (6), 2167–2195.

Battigalli, Pierpaolo, "On rationalizability in extensive games," *Journal of Economic Theory*, 1997, *74* (1), 40–61.

Bechtel, William, *Discovering cell mechanisms: The creation of modern cell biology*, Cambridge University Press, 2006.

Benassy, Jean-Pascal, "Chapter 4 Non-Walrasian equilibria, money, and macroeconomics," in "in," Vol. 1 of *Handbook of Monetary Economics*, Elsevier, 1990, pp. 103 – 169.

_ , "Nonclearing Markets: Microeconomic Concepts and Macroeconomic Applications," *Journal of Economic Literature*, 1993, *31* (2), 732–61.

Bentzel, Ragnar and Bent Hansen, "On Recursiveness and Interdependency in Economic Models," *Review of Economic Studies*, 1954, *22* (3), 153–168.

_ **and Herman Wold**, "On statistical demand analysis from the viewpoint of simultaneous equations," *Scandinavian Actuarial Journal*, 1946, *1946* (1), 95–114.

Bernheim, B. Douglas, "Rationalizable Strategic Behavior," *Econometrica*, 1984, *52* (4), 1007–1028.

Bianchi, Francesco, Cosmin Ilut, and Hikaru Saijo, "Diagnostic business cycles," *Review of Economic Studies*, 2024, *91* (1), 129–162.

Binmore, Ken, "Modeling rational players: Part I," *Economics & Philosophy*, 1987, *3* (2), 179–214.

Blanchard, Olivier and Jordi Galí, "Real wage rigidities and the New Keynesian model," *Journal of money, credit and banking*, 2007, *39*, 35–65.

Bollen, Kenneth A, "Total, direct, and indirect effects in structural equation models," *Sociological methodology*, 1987, pp. 37–69.

Bongers, Stephan, Tineke Blom, and Joris M. Mooij, "Causal Modeling of Dynamical Systems," *arXiv preprint arXiv:1803.08784*, 2022.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Diagnostic expectations and credit cycles," *The Journal of Finance*, 2018, *73* (1), 199–227.

Campbell, John Y, "Inspecting the mechanism: An analytical approach to the stochastic growth model," *Journal of Monetary Economics*, 1994, *33* (3), 463–506.

Carlson, Mark, Stefania D'Amico, Cristina Fuentes-Albero, Bernd Schlusche, and Paul R Wood, "Issues in the Use of the Balance Sheet Tool," Technical Report, Board of Governors of the Federal Reserve System (US) 2020.

Christiano, Lawrence J, Martin Eichenbaum, and Charles L Evans, "Nominal rigidities and the dynamic effects of a shock to monetary policy," *Journal of political Economy*, 2005, *113* (1), 1–45.

Citanna, Alessandro, Hervé Crès, Jacques Drèze, P.Jean-Jacques Herings, and Antonio Villanacci, "Continua of underemployment equilibria reflecting coordination failures, also at Walrasian prices," *Journal of Mathematical Economics*, 2001, *36* (3), 169 – 200.

Clower, Robert W, "The Keynesian counterrevolution: a theoretical appraisal," in Frank PR Brechling and Frank Hahn, eds., *The Theory of Interest Rates: Proceedings of a Conference*, London: Macmillan, 1965.

Craver, Carl F., *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*, Clarendon Press, 2007.

Crawford, Vincent P., "Let's talk it over: Coordination via preplay communication with level-k thinking," *Research in Economics*, 2017, *71* (1), 20–31.

_ , Miguel A. Costa-Gomes, and Nagore Iriberri, "Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications," *Journal of Economic Literature*, 2013, *51* (1), 5–62.

Darden, Lindley, "Mechanisms versus causes in biology and medicine," in "Mechanism and causality in biology and economics," Springer, 2013, pp. 19–34.

Davies, Dan, "why i am (still after all) an economist," 2023. https://backofmind.substack.com/p/why-i-am-still-after-all-an-economist, accessed on Feb 23rd, 2024.

De Regt, Henk W., *Understanding Scientific Understanding*, Oxford University Press, 2017.

_ **and Dennis Dieks**, "A contextual approach to scientific understanding," *Synthese*, 2005, *144*, 137–170.

Deaton, Angus, "Understanding the Mechanisms of Economic Development," *Journal of Economic Perspectives*, September 2010, *24* (3), 3–16.

Di Maggio, Marco, Amir Kermani, Benjamin J. Keys, Tomasz Piskorski, Rodney Ramcharan, Amit Seru, and Vincent Yao, "Interest Rate Pass-Through: Mortgage Rates, Household Consumption, and Voluntary Deleveraging," *The American Economic Review*, 2017, *107* (11), 3550–3588.

Drazen, Allan, "Recent Developments in Macroeconomic Disequilibrium Theory," *Econometrica*, March 1980, *48* (2), 283–306.

Dubey, Pradeep, "Price-Quantity Strategic Market Games," *Econometrica*, 1982, *50* (1), 111–126.

_ **, John Geanakoplos, and Martin Shubik**, "The revelation of information in strategic market games: A critique of rational expectations equilibrium," *Journal of Mathematical Economics*, 1987, *16* (2), 105–137.

Eusepi, Stefano and Bruce Preston, "A Short History in Defence of Adaptive Learning CAMA Working Paper 52/2023 October 2023," 2023.

Evans, George W. and Seppo Honkapohja, *Learning and expectations in macroeconomics*, Princeton University Press, 2001.

Farhi, Emmanuel and Iván Werning, "Monetary policy, bounded rationality, and incomplete markets," *American Economic Review*, 2019, *109* (11), 3887–3928.

Feldstein, Martin, "Commentary : Is there a role for discretionary fiscal policy?," *Proceedings - Economic Policy Symposium - Jackson Hole*, 2002, pp. 151–162.

Fisher, Franklin M., "A correspondence principle for simultaneous equation models," *Econometrica: Journal of the Econometric Society*, 1970, pp. 73–92.

_ , *Disequilibrium foundations of equilibrium economics*, Cambridge University Press, 1983.

Gabaix, Xavier, "A behavioral New Keynesian model," *American Economic Review*, 2020, *110* (8), 2271–2327.

García-Schmidt, Mariana and Michael Woodford, "Are low interest rates deflationary? A paradox of perfect-foresight analysis," *American Economic Review*, 2019, *109* (1), 86–120.

Gintis, Herbert, "The Dynamics of General Equilibrium," *Economic Journal*, October 2007, *117* (523), 1280–1309.

Giraud, Gael, "Strategic market games: an introduction," *Journal of Mathematical Economics*, 2003, *39* (5–6), 355 – 375. Strategic Market Games.

Glennan, Stuart S., "Mechanisms and the Nature of Causation," *Erkenntnis (1975-)*, 1996, *44* (1), 49–71.

Godley, Wynne and Marc Lavoie, *Monetary economics: an integrated approach to credit, money, income, production and wealth*, Springer, 2016.

Grandmont, Jean-Michel, "Temporary General Equilibrium Theory," *Econometrica*, April 1977, *45* (4), 535–572.

Grimm, S.R., C. Baumberger, and S. Ammon, *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, Taylor & Francis, 2016.

Grossman, Sanford J and Joseph E Stiglitz, "On the impossibility of informationally efficient markets," *The American economic review*, 1980, *70* (3), 393–408.

Guesnerie, Roger, "An exploration of the eductive justifications of the rational-expectations hypothesis," *The American Economic Review*, 1992, pp. 1254–1278.

Hall, Robert E and Ricardo Reis, "Achieving price stability by manipulating the central bank's payment on reserves," Technical Report, National Bureau of Economic Research 2016.

Hansson, Björn A., *The Stockholm School and the Development of Dynamic Method*, Croom Helm London, 1982.

Hart, Sergiu and Andreu Mas-Colell, "Uncoupled Dynamics Do Not Lead to Nash Equilibrium," *American Economic Review*, December 2003, *93* (5), 1830–1836.

Hayek, Friedrich A., "The Use of Knowledge in Society," *The American Economic Review*, 1945, *35* (4), 519–530.

Hazell, Jonathon, Juan Herreno, Emi Nakamura, and Jón Steinsson, "The slope of the Phillips Curve: evidence from US states," *The Quarterly Journal of Economics*, 2022, *137* (3), 1299–1344.

Heckman, James J and Rodrigo Pinto, "Econometric Causality: The Central Role of Thought Experiments," Working Paper 31945, National Bureau of Economic Research December 2023.

Hedström, Peter and Petri Ylikoski, "Causal mechanisms in the social sciences," *Annual Review of Sociology*, 2010, *36*, 49–67.

Heller, Walter Perrin and Ross M. Starr, "Unemployment Equilibrium with Myopic Complete Information," *The Review of Economic Studies*, 1979, *46* (2), 339–359.

Hills, Alison, "Understanding why," *Noûs*, 2016, *50* (4), 661–688.

Holden, Tom D., "Robust real rate rules," 2023.

Holland, Paul W, "Statistics and causal inference," *Journal of the American Statistical Association*, 1986, *81* (396), 945–960.

Howitt, Peter, "Interest rate control and nonconvergence to rational expectations," *Journal of Political Economy*, 1992, *100* (4), 776–800.

_ **and Robert Clower**, "The emergence of economic organization," *Journal of Economic Behavior & Organization*, January 2000, *41* (1), 55–84.

Imbens, Guido W, "Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics," *Journal of Economic Literature*, 2020, *58* (4), 1129–1179.

Iovino, Luigi and Dmitriy Sergeyev, "Central bank balance sheet policies without rational expectations," *Review of Economic Studies*, 2023, *90* (6), 3119–3152.

Iwai, Katsuhito, *Disequilibrium Dynamics: A Theoretical Analysis of Inflation and Unemployment*, Yale University Press, 1981.

_ , "Disequilibrium Dynamics of the Monetary Economy: A Micro-Founded Synthesis of the Wicksellian Theory of Cumulative Process and the Keynesian Theory of Effective Demand," *Available at SSRN 3178105*, 2019.

Jackson, Matthew O., "A crash course in implementation theory," *Social choice and welfare*, 2001, *18*, 655–708.

Jordan, James S., "A dynamic model of expectations equilibrium," *Journal of Economic Theory*, 1982, *28* (2), 235–254.

Judd, Charles M. and David A. Kenny, "Process analysis: Estimating mediation in treatment evaluations," *Evaluation Review*, 1981, *5* (5), 602–619.

Kaplan, Greg, Benjamin Moll, and Giovanni L Violante, "Monetary policy according to HANK," *American Economic Review*, 2018, *108* (3), 697–743.

Keane, Michael, "Structural vs. atheoretic approaches to econometrics," *Journal of Econometrics*, 2010, *156* (1), 3–20.

Kehoe, Timothy, "An evaluation of the performance of applied general equilibrium models of the impact of NAFTA," in Timothy J. Kehoe, T. N. Srinivasan, and John Whalley, eds., *Frontiers in applied general equilibrium modeling: In honor of Herbert Scarf*, Cambridge University Press, 2005.

_ **, Pau Pujolas, and Jack Rossbach**, "Quantitative Trade Models: Developments and Challenges," *Annual Review of Economics*, 2017, *9* (1), 295–325.

Keynes, John M., *The General Theory of Employment, Interest and Money*, London: Macmillan, 1936.

Kocherlakota, Narayana R., "Bounds on price-setting," *Theoretical Economics*, 2021, *16* (3), 979–1015.

Kreps, David M., "A note on "fulfilled expectations" equilibria," *Journal of Economic Theory*, 1977, *14* (1), 32–43.

_ , "Anticipated Utility and Dynamic Choice," in Ehud Kalai Donald P. Jacobs and Morton I. Kamien, eds., *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures, 1983–1997*, Cambridge University Press, 1998, p. 242–274.

Lange, Marc, *An introduction to the philosophy of physics: Locality, fields, energy, and mass*, Blackwell Publishing, 2002.

Leijonhufvud, Axel, *On Keynesian Economics and the Economics of Keynes: A Study in Monetary Theory*, Oxford University Press, 1968.

_ , "Effective Demand Failures," *The Swedish Journal of Economics*, 1973, *75* (1), 27–48.

LeRoy, Stephen F, *Causal inference in economic models*, Cambridge Scholars Publishing, 2020.

Lindahl, Erik, *Studies in the Theory of Money and Capital*, George Allen & Unwin Ltd, 1939.

Lindh, Thomas, "Lessons from Learning to Have Rational Expectations," Working Paper Series, Research Institute of Industrial Economics December 1989.

Lucas, Robert E., "Econometric policy evaluation: A critique," *Carnegie-Rochester Conference Series on Public Policy*, 1976, *1*, 19 – 46.

_ , "Methods and Problems in Business Cycle Theory," *Journal of Money, Credit and Banking*, 1980, *12* (4), 696–715.

Lundberg, Erik, *Studies in the theory of economic expansion*, P.S. King & Son, ltd., 1937.

Machamer, Peter, Lindley Darden, and Carl F. Craver, "Thinking about Mechanisms," *Philosophy of Science*, 03 2000, *67.*

MacKinnon, David, *Introduction to statistical mediation analysis*, Routledge, 2012.

Mandel, Antoine and Herbert Gintis, "Decentralized pricing and strategic stability of Walrasian general equilibrium," *J. Math. Econ*, 2016, *63*, 84–92.

Marcet, Albert and Thomas Sargent, "The convergence of Vector Autoregressions to rational expectations equilibria," in Alessandro Vercelli and Nicola Dimitri, eds., *Macroeconomics: A Survey of Research Strategies*, Oxford University Press, 1992, pp. 139–164.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green, *Microeconomic Theory*, Oxford University Press, 1995.

Matějka, Filip and Alisdair McKay, "Rational inattention to discrete choices: A new foundation for the multinomial logit model," *American Economic Review*, 2015, *105* (1), 272–298.

Mishkin, Frederic S., "Symposium on the monetary transmission mechanism," *Journal of Economic perspectives*, 1995, *9* (4), 3–10.

Morgan, Mary S., "The stamping out of process analysis in econometrics," in Neil De Marchi and Mark Blaug, eds., *Appraising economic theories*, Edward Elgar Publishing, 1991.

Nash, John F., "Equilibrium points in n-person games," *Proceedings of the national academy of sciences*, 1950, *36* (1), 48–49.

Pálvölgyi, Dömötör, Gyuri Venter, and Liyan Yang, "Multiple Equilibria in Noisy Rational Expectations Economies," 2017.

Pawson, Ray and Nick Tilley, *Realistic evaluation*, SAGE, 1997.

Pearce, David G., "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 1984, *52* (4), 1029–1050.

Pearl, Judea, "Direct and Indirect Effects," in "Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence" UAI'01 Morgan Kaufmann Publishers Inc. San Francisco, CA, USA 2001, p. 411–420.

_ , *Causality*, Cambridge University Press, 2009.

_ **and Dana Mackenzie**, *The Book of Why: The New Science of Cause and Effect*, Penguin Books, Limited, 2018.

Peters, Craig, "Evaluating the Performance of Merger Simulation: Evidence from the U.S. Airline Industry," *The Journal of Law and Economics*, 2006, *49* (2), 627–649.

Ravenna, Federico and Carl E Walsh, "Optimal monetary policy with the cost channel," *Journal of Monetary Economics*, 2006, *53* (2), 199–216.

Richardson, Thomas S., "Models of feedback: interpretation and discovery." PhD dissertation, Carnegie-Mellon University 1996.

Roberts, John, "An Equilibrium Model with Involuntary Unemployment at Flexible, Competitive Prices and Wages," *American Economic Review*, 1987, *77* (5), 856–874.

_ , "Involuntary unemployment and imperfect competition: a game-theoretic macromodel," in "The economics of imperfect competition and employment: Joan Robinson and beyond," Springer, 1989, pp. 146–165.

Robins, James M. and Sander Greenland, "Identifiability and exchangeability for direct and indirect effects," *Epidemiology*, 1992, *3* (2), 143–155.

Rubin, Donald B., "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, 1974, *66* (5), 688.

Rudd, Jeremy B, "Why do we think that inflation expectations matter for inflation?(And should we?)," *Review of Keynesian Economics*, 2022, *10* (1), 25–45.

Rupert, Peter and Roman Sustek, "On the Mechanics of New-Keynesian Models," *Journal of Monetary Economics*, 2019.

Saari, Donald G. and Carl P. Simon, "Effective Price Mechanisms," *Econometrica*, 1978, *46* (5), 1097–1125.

Samuelson, Paul A., "A Synthesis of the Principle of Acceleration and the Multiplier," *Journal of Political Economy*, 1939, *47* (6), 786–797.

Scarf, Herbert, "Some Examples of Global Instability of the Competitive Equilibrium," *International Economic Review*, 1960, *1* (3).

Schinkel, Maarten Pieter, *Disequilibrium Theory: Reflections Towards a Revival of Learning*, Universitaire Pers Maastricht, 2001.

Shapley, Lloyd and Martin Shubik, "Trade Using One Commodity as a Means of Payment," *The Journal of Political Economy*, 1977, *85* (5), 937–968.

Simon, Herbert, "Causal Ordering and Identifiability," in Wm. C. Hood and Tjalling C. Koopmans, eds., *Studies in Econometric Method*, 1953.

Sims, Christopher A, "Implications of rational inattention," *Journal of monetary Economics*, 2003, *50* (3), 665–690.

Strotz, Robert H. and Herman O. A. Wold, "Recursive vs. Nonrecursive Systems: An Attempt at Synthesis (Part I of a Triptych on Causal Chain Systems)," *Econometrica*, 1960, *28* (2), 417–427.

Tesfatsion, Leigh, "Modeling economic systems as locally-constructive sequential games," *Journal of Economic Methodology*, 2017, *24* (4), 384–409.

Thagard, Paul, *How scientists explain disease*, Princeton University Press, 2000.

VanderWeele, Tyler, *Explanation in causal inference: methods for mediation and interaction*, Oxford University Press, 2015.

Werning, Iván, "Expectations and the Rate of Inflation," Technical Report, National Bureau of Economic Research 2022.

Wold, Herman O. A., "Causality and Econometrics," *Econometrica*, Apr 1954, *22* (2), 162–177.

Woodford, Michael, "Control of the Public Debt: A Requirement for Price Stability?," Working Paper 5684, National Bureau of Economic Research July 1996.

_ , "Macroeconomic analysis without the rational expectations hypothesis," *Annual Review of Economics*, 2013, *5* (1), 303–346.

Woodward, James, *Making Things Happen: A Theory of Causal Explanation (Oxford Studies in the Philosophy of Science)*, Oxford University Press, USA, 2003.

Young, H Peyton, *Strategic learning and its limits*, OUP Oxford, 2004.

A Proofs

A.1 Decomposition in the baseline RANK model

Household optimization yields the Euler equation and labor supply optimality conditions:

$$\begin{aligned} c_t^{-\gamma} &= \beta(1+r_t)c_{t+1}^{-\gamma} \\ w_t c_t^{-\gamma} &= \varphi n_t^{\nu} \end{aligned}$$

To calculate the direct effect of a change in interest rates, we solve forward the household budget constraint assuming w and T are constant at their steady state levels. Using the transversality and no-Ponzi conditions,

$$\begin{aligned} \sum_{t=0}^{\infty}\prod_{k=0}^{t-1}(1+r_k)^{-1}c_t &= \sum_{t=0}^{\infty}\prod_{k=0}^{t-1}(1+r_k)^{-1}(wn_t+T) \\ \sum_{t=0}^{\infty}\prod_{k=0}^{t-1}(1+r_k)^{-1}(\beta(1+r_k))^{\frac{1}{\gamma}}c_0 &= \sum_{t=0}^{\infty}\prod_{k=0}^{t-1}(1+r_k)^{-1}(\beta(1+r_k))^{-\frac{1}{\nu}}c_0^{-\frac{\gamma}{\nu}}w^{\frac{1+\nu}{\nu}}\varphi^{-\frac{1}{\nu}} + \sum_{t=0}^{\infty}\prod_{k=0}^{t-1}(1+r_k)^{-1}T \end{aligned}$$

Differentiating with respect to c_0 and r_0 and evaluating at steady state,

$$\frac{dc_0}{1-\beta}+\left(\frac{1}{\gamma}-1\right)\frac{c_0\beta^2}{1-\beta}dr_0=-\frac{\gamma}{\nu}\frac{c_0^{-\frac{\gamma}{\nu}-1}w^{\frac{1+\nu}{\nu}}\varphi^{-\frac{1}{\nu}}}{1-\beta}dc_0-\frac{1+\nu}{\nu}\frac{c_0^{-\frac{\gamma}{\nu}}w^{\frac{1+\nu}{\nu}}\varphi^{-\frac{1}{\nu}}\beta^2}{1-\beta}dR_0-\frac{T\beta^2}{1-\beta}dr_0$$

$$\left(1+\frac{\gamma}{\nu}\frac{wn}{c_0}\right)dc_0=-\frac{1}{\gamma}\left(1+\frac{\gamma}{\nu}\frac{wn}{c_0}\right)c_0\beta^2dr_0$$

where we use the fact that $c_0=wn+T$ in steady state. This implies that the direct effect is $dc_0=-\frac{1}{\gamma}c_0\beta^2dr_0$. Since the total effect is $dc_0=-\frac{1}{\gamma}\beta c_0dr_0$ (from the aggregate Euler equation, using the fact that the economy returns to steady state at date 1 in general equilibrium) we are done.

A.2 Real rate rule and working capital economy

Under a **real rate rule**, we have

$$\begin{aligned}y_t&=y_{t+1}-\gamma^{-1}(i_t-\pi_{t+1})\\\pi_t&=\kappa y_t+\beta\pi_{t+1}\\i_t&=r_t+\phi\pi_t+\varepsilon_t\\i_t&=r_t+\pi_{t+1}\end{aligned}$$

where with some abuse of notation we let r_t denote the log-deviation of $1+r_t$. Combining the last two equations yields $\pi_{t+1}=\phi\pi_t+\varepsilon_t$. Since $\phi>1$, the unique bounded solution is $\pi_0=-\frac{1}{\phi}\varepsilon_0$, $\pi_t=0$ for all $t>0$. Substituting into the Phillips curve, $y_t=\pi_t/\kappa$ implying $y_0=-\frac{1}{\kappa\phi}\varepsilon_0$, $y_t=0$ for $t>0$. Substituting into the Euler equation, $i_0=r_0=\frac{\gamma}{\kappa\phi}\varepsilon_0$, as claimed in the main text. Since the household problem is unchanged, the direct-indirect effect decomposition is as before.

In the **working capital economy**, the representative firm has technology $Y_t(i)=n_t(i)^\alpha$ and faces marginal cost $wY_t(i)^{1/\alpha}(1+i_t)$, given that real wages are constant at w. Its problem is

$$\max_{\{P_t(i),Y_t(i)\}_{t=0}^\infty}\sum_{t=0}^\infty Q_{t|0}\left[\frac{P_t(i)}{P_t}-wY_t(i)^{\frac{1}{\alpha}}(1+i_t)-\frac{\psi}{2}Y_t\left(\frac{P_t(i)}{P_{t-1}(i)}-1\right)^2\right]$$

$$\text{s.t. } Y_t\left(\frac{P_t(i)}{P_t}\right)^{-\varepsilon}=Y_t(i)$$

where $Q_{t|0}=\prod_{k=0}^{t-1}\left(\frac{1}{1+r_k}\right)$. Taking first order conditions and imposing a symmetric equilibrium yields the Phillips curve

$$\Pi_t(\Pi_t-1)=\frac{\varepsilon}{\psi}\left[\frac{w}{\alpha}Y_t^{\frac{1-\alpha}{\alpha}}(1+i_t)-\frac{\varepsilon-1}{\varepsilon}\right]+\frac{1}{1+r_t}\frac{Y_{t+1}}{Y_t}\Pi_{t+1}(\Pi_{t+1}-1)$$

Given the real wage w, output in the zero inflation steady state is $Y=\left(\frac{\alpha}{\beta w}\frac{\varepsilon-1}{\varepsilon}\right)^{\frac{\alpha}{1-\alpha}}$. Log-linearizing

around this steady state, we have the expression in the main text,

$$\pi_t = \kappa y_t + \frac{\kappa\alpha}{1-\alpha} i_t + \beta \pi_{t+1}$$

where $\kappa = \frac{\varepsilon - 1}{\psi} \frac{1-\alpha}{\alpha}$. Assuming a one-time shock to the standard Taylor rule (so the economy returns to steady state at date 1) and using the Euler equation $i_0 = -\gamma y_0$ to substitute out for i_t, this becomes $\pi_0 = \kappa \left(1 - \frac{\gamma\alpha}{1-\alpha}\right) y_0$. Using both equations in the Taylor rule:

$$-\gamma y_0 = \phi\kappa \left(1 - \frac{\gamma\alpha}{1-\alpha}\right) y_0 + \varepsilon_0$$
$$y_0 = -\frac{1}{\phi\kappa + \gamma\left(1 - \frac{\phi\kappa\alpha}{1-\alpha}\right)} \varepsilon_0$$

as in the main text. When $\frac{\phi\kappa\alpha}{1-\alpha} = 1$, $\frac{dy_0}{d\varepsilon_0} = -\frac{1}{\phi\kappa}$; when $\frac{\phi\kappa\alpha}{1-\alpha} > 1$, a higher IES (lower γ) reduces $\left|\frac{dy_0}{d\varepsilon_0}\right|$. Finally, to compute the direct effect of monetary policy on household consumption in the working capital economy, suppose real interest rates change but all other variables affecting household decisions remain unchanged. In that case, household income remains unchanged at its steady state value Y, and the household lifetime budget constraint can be written

$$\sum_{t=0}^{\infty} \prod_{k=0}^{t-1} (1+r_k)^{-1} c_t = \sum_{t=0}^{\infty} \prod_{k=0}^{t-1} (1+r_k)^{-1} Y$$
$$\sum_{t=0}^{\infty} \prod_{k=0}^{t-1} (1+r_k)^{-1} (\beta(1+r_k))^{\frac{1}{\gamma}} c_0 = \sum_{t=0}^{\infty} \prod_{k=0}^{t-1} (1+r_k)^{-1} Y$$

Differentiating with respect to c_0 and r_0 and evaluating at steady state, the direct effect is $dc_0 = -\frac{1}{\gamma} Y \beta^2 dR_0$, as in our baseline model. Since the total effect is also the same as in our baseline model, it follows that the share of direct and indirect effects is the same.

Introducing HtM households and government debt A fraction $\eta \in (0,1)$ of households are hand to mouth, and cannot trade assets. The remaining $1-\eta$ households are unconstrained. All households pay per capita real lump sum taxes $\mathcal{T}_t$. For simplicity, we assume as in the working capital economy that real wages are rigid. We also assume that firms demand the same quantity of labor from HtMs and unconstrained households, and all households receive an equal share of profits. Thus, an unconstrained household's income (and therefore their consumption) is $C_t^u = Y_t - \mathcal{T}_t$.

The government issues one-period nominal debt: the government budget constraint is

$$\frac{B_{t+1}}{1+i_t} = B_t - P_t \mathcal{T}_t$$

We assume that the government adjusts $\mathcal{T}_t$ as necessary to keep $\bar{b} := \frac{B_{t+1}}{P_t} > 0$ constant. Dividing

by P_t, we have

$$\frac{\bar{b}}{1+i_t} = \bar{b}\Pi_t - \mathcal{T}_t$$

where $\Pi_t := P_t/P_{t-1}$. In the zero-inflation steady state, $\mathcal{T} = (1-\beta)\bar{b}$. Log-linearizing around this steady state,

$$\tau_t = \frac{\beta}{1-\beta} i_t - \frac{1}{1-\beta}\pi_t$$

Log-linearizing the aggregate resource constraint,

$$\begin{aligned} Y_t &= \eta C_t^h + (1-\eta) C_t^u = \eta Y_t - \eta \mathcal{T}_t + (1-\eta) C_t^u \\ y_t &= \eta y_t - \eta\vartheta\tau_t + (1-\eta+\eta\vartheta) c_t^u \end{aligned}$$

where C_t^u denotes unconstrained households' consumption, c_t^u its log deviation, $\vartheta := \mathcal{T}/y$ denotes the steady state tax-to-GDP ratio, and we use the fact that $C^u = Y + \frac{\eta}{1-\eta}\mathcal{T}$ in steady state. As before, unconstrained houseolds' consumption satisfies an Euler equation, which we can write in log-linear form

$$c_t^u = c_{t+1}^u - \frac{1}{\gamma}(i_t - \pi_{t+1})$$

By construction, all variables return to steady state from date 1 onwards following a one-off shock. Thus, the system with a real rate rule can be written

$$\begin{aligned} i_0 &= r_0 + \phi\pi_0 + \varepsilon_0 \\ r_0 &= i_0 \\ \pi_0 &= \kappa y_0 \\ y_0 &= \eta y_0 - \eta\vartheta\tau_0 + (1-\eta+\eta\vartheta) c_0^u \\ c_t^u &= -\frac{1}{\gamma} i_0 \\ \tau_0 &= \frac{\beta}{1-\beta} i_0 - \frac{1}{1-\beta}\pi_0 \end{aligned}$$

As before, the first three equations can be solved for $\pi_0 = -\frac{1}{\phi}\varepsilon_0$, $y_0 = -\frac{1}{\kappa\phi}$. Substituting this into the remaining three equations yields

$$\begin{aligned} \tau_0 &= \frac{\beta}{1-\beta} i_0 + \frac{1}{1-\beta}\frac{1}{\phi}\varepsilon_0 \\ -(1-\eta)\frac{1}{\kappa\phi} &= -\eta\vartheta\tau_0 - (1-\eta+\eta\vartheta)\frac{1}{\gamma} i_0 \end{aligned}$$

which can be solved to yield

$$i_0 = \frac{(1-\eta)(1-\beta)\kappa^{-1} - \eta\vartheta}{\eta\vartheta\beta + [1-\eta+\eta\vartheta]\gamma^{-1}(1-\beta)}\frac{\varepsilon_0}{\phi}, \qquad \tau_0 = \left[\frac{(1-\eta)\beta\kappa^{-1} + [1-\eta+\eta\vartheta]\gamma^{-1}}{\eta\vartheta\beta + [1-\eta+\eta\vartheta]\gamma^{-1}(1-\beta)}\right]\frac{\varepsilon_0}{\phi}$$

To compute the indirect effect of monetary policy on aggregate consumption via taxes, consider a change $d\mathcal{T}_0$ in real taxes while interest rates and incomes remain unchanged. Using their lifetime budget constraint, unconstrained households' consumption changes by $dC_0^u = (1-\beta)d\mathcal{T}_0$, while HtMs adjust consumption by $d\mathcal{T}_0$ (their marginal propensity to consume is 1). Since the total effect is $dY_0 = -Y_0(1/\phi\kappa)\varepsilon_0$, the share due to indirect effects via taxes is

$$\frac{[\eta + (1-\eta)(1-\beta)]d\mathcal{T}_0}{dY_0} = [\eta + (1-\eta)(1-\beta)]\vartheta \frac{(1-\eta)\beta + [1-\eta+\eta\vartheta]\gamma^{-1}\kappa}{\eta\vartheta\beta + [1-\eta+\eta\vartheta]\gamma^{-1}(1-\beta)}$$

which is increasing in ϑ (equivalently, in the debt-to-GDP ratio $\frac{B}{PY} = \frac{\vartheta}{1-\beta}$). Yet, as claimed in the main text, the aggregate consumption response $y_0 = -\frac{1}{\phi\kappa}$ does not depend on this ratio.

A.3 Defining direct effects 'even more directly'

While Kaplan et al. (2018) define the direct effect of monetary policy in terms of the change in real interest rates, they note that one could define it 'even more directly' in terms of the shock ε_t, by substituting the Taylor rule into the household's problem. In our setting, this amounts to

$$\begin{aligned} y_0 &= -\frac{\beta}{\gamma} i_0 + (1-\beta) y_0 \\ &= \underbrace{-\frac{\beta}{\gamma}\varepsilon_0}_{\text{direct effect}} \underbrace{-\frac{\beta\phi}{\gamma}\pi_0 + (1-\beta)y_0}_{\text{indirect effects}} \end{aligned}$$

where now 'indirect effects' include the effect of the change in real interest rates induced by the monetary authority's endogenous reaction to inflation. This does not rule out the paradox: the IES may still be irrelevant for the response of y_0 to ε_0 whatever the share of 'direct effects'. In the model with working capital and $\frac{\phi\kappa\alpha}{1-\alpha} = 1$, $\pi_0 = \kappa\left(1 - \frac{\gamma\alpha}{1-\alpha}\right) y_0 = -\frac{1}{\phi}\left(1 - \frac{\gamma\alpha}{1-\alpha}\right)\varepsilon_0$. So the decomposition becomes

$$-\frac{1}{\phi\kappa}\varepsilon_0 = \underbrace{-\frac{\beta}{\gamma}\varepsilon_0}_{\text{direct effect}} + \underbrace{\left(\frac{1}{\gamma} - \frac{\alpha}{1-\alpha}\right)\varepsilon_0 - \frac{1-\beta}{\phi\kappa}\varepsilon_0}_{\text{indirect effects}}$$

The share of direct effects is now $\frac{\beta\phi\kappa}{\gamma}$, which can be less than, equal to, or even greater than 1. Even if this share is close to 1 – i.e. the last two terms approximately sum to zero – a change in γ^{-1} does not effect $dy_0/d\varepsilon_0$. While it changes the magnitude of the direct effect, it also changes the size of indirect effects, namely the effect of the equilibrium response of inflation, via the Taylor rule, on the interest rate faced by households. In fact, these two changes exactly offset each other.

A.4 Proposition 3.1

If (5) is recursive, $\mathbf{A}$ is a triangular matrix with ones on the diagonal, and therefore invertible, so the system has unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}z$, $y = \mathbf{g}^\intercal\mathbf{x} = \mathbf{g}^\intercal\mathbf{A}^{-1}\mathbf{c}z$. Since the solution for $\mathbf{x}$ does not

depend on $\mathbf{g}$, we have

$$\frac{d^2y}{dg_i dz} = \frac{dx_i}{dz}$$
$$\epsilon_i = \frac{d^2y}{dg_i dz}\frac{g_i}{dy/dz} = \frac{dx_i}{dz}\frac{g_i}{dy/dz} = s_i$$

To prove claim (i), suppose (5) is unrestricted. We can write the system as

$$\underbrace{\begin{bmatrix} \mathbf{A} & \mathbf{b} \\ -\mathbf{g}^\intercal & 1 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} \mathbf{x} \\ y \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ 0 \end{bmatrix} z$$

This system has a unique solution iff the blocked matrix $\mathbf{B}$ is invertible. Since its bottom right block [1] is invertible, $\mathbf{B}$ is invertible if the Schur complement of [1], namely $\mathbf{A} + \mathbf{b}\mathbf{g}^\intercal$, is invertible. The solution is then $\mathbf{x} = (\mathbf{A} + \mathbf{b}\mathbf{g}^\intercal)^{-1}\mathbf{c}z$, $y = \mathbf{g}^\intercal(\mathbf{A} + \mathbf{b}\mathbf{g}^\intercal)^{-1}\mathbf{c}z$. Note that if $\mathbf{b} = \mathbf{0}$ (y does not appear in the first $n-1$ equations) then $\mathbf{x}$ does not depend on $\mathbf{g}$, as in the recursive case.

Substituting the last equation of (5) into the first $n-1$ equations,

$$\begin{aligned}
\mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{g}^\intercal\mathbf{x} &= \mathbf{c}z \\
(\mathbf{A} + \mathbf{b}\mathbf{g}^\intercal)\frac{d\mathbf{x}}{dz} &= \mathbf{c} \\
(\mathbf{A} + \mathbf{b}\mathbf{g}^\intercal)\frac{d^2\mathbf{x}}{dg_i dz} + \mathbf{b}\frac{dx_i}{dz} &= \mathbf{0} \\
\frac{d^2\mathbf{x}}{dg_i dz} &= -(\mathbf{A} + \mathbf{b}\mathbf{g}^\intercal)^{-1}\mathbf{b}\frac{dx_i}{dz} \\
\frac{d^2y_i}{dg_i dz} &= \frac{dx_i}{dz} + \mathbf{g}^\intercal\frac{d^2\mathbf{x}}{dg_i dz} \\
&= \left(1 - \mathbf{g}^\intercal(\mathbf{A} + \mathbf{b}\mathbf{g}^\intercal)^{-1}\mathbf{b}\right)\frac{dx_i}{dz} \\
\epsilon_i = \frac{d^2y}{dg_i dz}\frac{g_i}{dy/dz} &= \left(1 - \mathbf{g}^\intercal(\mathbf{A} + \mathbf{b}\mathbf{g}^\intercal)^{-1}\mathbf{b}\right)s_i
\end{aligned}$$

Clearly there exist cases where the term in brackets is negative, implying that ϵ_i and s_i have opposite sign. For example, take $\mathbf{A} = \mathbf{I}_{n-1} - \mathbf{b}\mathbf{g}^\intercal$, $\mathbf{g}^\intercal\mathbf{b} > 1$.

To prove claim (ii), suppose $\mathbf{b}$ is not contained in the column space of $\mathbf{A}$, i.e. there does not exist $\mathbf{w} \in \mathbb{R}^{n-1}$ such that $\mathbf{A}\mathbf{w} = \mathbf{b}$ (which implies $\mathbf{A}$ has rank less than $n-1$). Consider a class of systems (5) which differ only in g_i for some i; call their solutions $(\mathbf{x}(g_i), y(g_i))$. We will show that $y(g_i)$ does not depend on g_i. Fix z and take two different values of g_i, say $g_i^0 \neq g_i^1$. We must have

$$\begin{aligned}
\mathbf{A}\mathbf{x}(g_i^0) + \mathbf{b}y(g_i^0) &= \mathbf{c}z \\
\mathbf{A}\mathbf{x}(g_i^1) + \mathbf{b}y(g_i^1) &= \mathbf{c}z \\
\mathbf{A}(\mathbf{x}(g_i^1) - \mathbf{x}(g_i^0)) + \mathbf{b}(y(g_i^1) - y(g_i^0)) &= 0
\end{aligned}$$

If $y(g_i^1) \neq y(g_i^0)$, then

$$\mathbf{w} := -\frac{\mathbf{x}(g_i^1) - \mathbf{x}(g_i^0)}{y(g_i^1) - y(g_i^0)}$$

is a vector satisfying $\mathbf{Aw} = \mathbf{b}$, a contradiction. This implies that $\epsilon_i = \frac{d^2y}{dg_i dz}\frac{g_i}{dy/dz} = 0$, assuming of course that $\frac{dy}{dz} \neq 0$ (if not, the decomposition is not well-defined). □

A.5 Proposition 3.2

In a recursive system, $\frac{dy}{d\widetilde{z}} = g_i \frac{dx_i}{dz}$, so by definition of s_i we have

$$s_i = g_i \frac{dx_i/dz}{dy/dz} = \frac{dy/d\widetilde{z}}{dy/dz}.$$

In an unrestricted system,

$$\begin{aligned}
\mathbf{Ax} + \mathbf{bg}^\intercal \mathbf{x} + \mathbf{b}h\widetilde{z} &= \mathbf{c}z \\
\frac{d\mathbf{x}}{d\widetilde{z}} &= -(\mathbf{A} + \mathbf{bg}^\intercal)^{-1}\mathbf{b}h \\
\frac{dy}{d\widetilde{z}} &= h + \mathbf{g}^\intercal \frac{d\mathbf{x}}{d\widetilde{z}} \\
&= (1 - (\mathbf{A} + \mathbf{bg}^\intercal)^{-1}\mathbf{b})h \\
\frac{dy/d\widetilde{z}}{dy/dz} &= (1 - (\mathbf{A} + \mathbf{bg}^\intercal)^{-1}\mathbf{b})g_i \frac{dx_i/dz}{dy/dz} \\
&= (1 - (\mathbf{A} + \mathbf{bg}^\intercal)^{-1}\mathbf{b})s_i
\end{aligned}$$

Again, there exist cases where the term in parentheses is negative, implying that $\frac{dy/d\widetilde{z}}{dy/dz}$ and s_i have opposite sign.

Suppose $\mathbf{b}$ is not contained in the column space of $\mathbf{A}$. If $(\mathbf{x}, y)$ solves our system for $\widetilde{z} = 0$, we must have $y = 0$. For, if $y \neq 0$, define $\mathbf{w} = -y^{-1}\mathbf{x}$, and we have $\mathbf{Aw} = \mathbf{b}$, a contradiction. Now let $(\mathbf{x}(\widetilde{z}), y(\widetilde{z}))$ denote the solution for any $\widetilde{z}$ and take $\widetilde{z}^0 \neq \widetilde{z}^1$. We have

$$\begin{aligned}
\mathbf{Ax}(\widetilde{z}^0) + \mathbf{b}y(\widetilde{z}^0) &= 0 \\
\mathbf{Ax}(\widetilde{z}^1) + \mathbf{b}y(\widetilde{z}^1) &= 0 \\
\mathbf{A}(\mathbf{x}(\widetilde{z}^1) - \mathbf{x}(\widetilde{z}^0)) + \mathbf{b}(y(\widetilde{z}^1) - y(\widetilde{z}^0)) &= 0
\end{aligned}$$

If $y(\widetilde{z}^1) \neq y(\widetilde{z}^0))$, then

$$\mathbf{w} := -\frac{\mathbf{x}(\widetilde{z}^1) - \mathbf{x}(\widetilde{z}^0)}{y(\widetilde{z}^1) - y(\widetilde{z}^0)}$$

is a vector satisfying $\mathbf{Aw} = \mathbf{b}$, a contradiction. So $y(\widetilde{z}^1) = y(\widetilde{z}^0))$ for all $\widetilde{z}^0, \widetilde{z}^1$, i.e. $dy/d\widetilde{z} = 0$. □

A.6 Proposition 4.1

Each farmer's optimality conditions are

$$c_t(i)^{-\sigma}\frac{p_t(i)}{p_t} = \frac{\gamma}{\gamma-1}v'\left(y_t\left(\frac{p_t(i)}{p_t}\right)^{-\gamma}\right), t = 0, ..., T \tag{18}$$

$$c_t(i)^{-\sigma} = \beta\frac{p_t}{p_{t+1}}\left(c_{t+1}(i)\right)^{-\sigma}, t = 0, ..., T-1 \tag{19}$$

Combining i's flow budget constraints, we have the lifetime budget constraint

$$\sum_{t=0}^{T} p_t c_t(i) = \sum_{t=0}^{T} p_t y_t \left(\frac{p_t(i)}{p_t}\right)^{1-\gamma}$$

(19) implies $c_t(i)^{-\sigma} = \frac{p_t}{\beta p_0}c_0(i)^{-\sigma}$. Substituting this into (18) and the lifetime budget constraint:

$$\frac{p_t}{\beta p_0}c_0(i)^{-\sigma}\frac{p_t(i)}{p_t} = \frac{\gamma}{\gamma-1}v'\left(y_t\left(\frac{p_t(i)}{p_t}\right)^{-\gamma}\right)$$

$$\sum_{t=0}^{T} p_t c_0(i)\left(\frac{p_t}{\beta p_0}\right)^{-1/\sigma} = \sum_{t=0}^{T} p_t y_t \left(\frac{p_t(i)}{p_t}\right)^{1-\gamma}$$

Taking aggregate variables as given, the first equation defines an increasing relation between $c_0(i)$ and $p_t(i)$ for any t, while the second defines a decreasing relation between $c_0(i)$ and $p_t(i)$ for each t (since $\gamma > 1$). It follows that $c_0(i)$ and $p_t(i)$ must each be the same for all i. (Suppose $c_0(i) > c_0(j)$ for some $i \neq j$: then the first equation implies $p_t(i) > p_t(j)$ for each t, and the second implies $c_0(i) < c_0(j)$, a contradiction.) Thus, the definition of the price index implies $p_t = p_t(i)$, which implies $y_t(i) = y_t = c_t = c_t(i)$. This in turn implies $m_{t+1}(i) = 0$ for all t and i. (18) becomes

$$y_t^{-\sigma} = \frac{\gamma}{\gamma-1}v'(y_t), t = 0, ..., T$$

or in other words $y_t = y^*$, as defined in the Proposition. Substituting into (19) yields $1 = \beta\frac{p_t}{p_{t+1}}$, as claimed. All equilibrium conditions are homogeneous of degree 0 in the whole price sequence $\{p_t\}_{t=0}^T$: if $\{p_t\}_{t=0}^T$ is an equlibrium, $\{\lambda p_t\}_{t=0}^T$ is an equilibrium for any $\lambda > 0$. So we are done. □

A.7 Proposition 4.2

Given the maintained assumption of equal initial credit balances $m_t(i) = 0$, the subgame beginning in the morning of period t is isomorphic to a full game with $T - t$ periods. Thus, the first part of the Proposition follows immediately from Proposition 4.1. To prove the second part, note first that any equilibrium of the date t afternoon subgame must satisfy all equilibrium conditions of the date $t + 1$ morning subgame. Thus, from the first part, we have $c_\tau(i) = c_\tau = y_\tau = y^*$, $p_\tau(i) = p_\tau$ and $p_{\tau+1} = \beta p_\tau$ for all $\tau > t$. Combining farmer i's flow budget constraints and using the fact that $m_{T+1}(i) = 0$ under an optimal plan, we can write his optimization problem in the date t afternoon

subgame as

$$\max_{\{c_\tau(i)\}_{\tau=t}^T, \{p_\tau(i)\}_{\tau=t+1}^T} \sum_{\tau=t}^T \beta^{T-t} \left[\frac{c_\tau(i)^{1-\sigma}}{1-\sigma} - v\left(y_\tau \left(\frac{p_\tau(i)}{p_\tau} \right)^{-\gamma} \right) \right]$$
$$\text{s.t.} \quad \sum_{\tau=t}^T p_\tau c_\tau(i) = \sum_{\tau=t}^T p_\tau(i) \left(\frac{p_\tau(i)}{p_\tau} \right)^{-\gamma} y_\tau$$

which yields the Euler equation

$$c_t(i)^{-\sigma} = \beta \frac{p_t}{p_{t+1}} c_{t+1}(i)^{-\sigma}$$

Since we know $c_{t+1}(i) = y^*$, this implies $y_t = c_t = c_t(i) = (p_{t+1}/(\beta p_t))^{-1/\sigma} y^*$ for all i. Provided that (y_t, p_t, p_{t+1}) satisfy this relation, the farmer's remaining optimality conditions are all satisfied. So we are done. □

A.8 Interest on reserves game

Integrating the household budget constraint $A_i = M + p - C_i + q - B_i$ across all traders and using the definition of prices p, q, we have $\int A_i di = M + \delta$. So lump sum taxes are $T = (1+r)\int A_i di - \delta/q = (1+r)(M+\delta) - \delta/q$. Using this definition and the budget constraint in (10), we can rewrite trader i's objective function:

$$\theta \ln\left(\frac{C_i}{p}\right) + \ln\left(\frac{B_i + \delta}{q} + (1+r)[p - C_i + B - B_i]\right) \tag{20}$$

In the subgame beginning in the afternoon of date 1, i chooses $B_i \in [0, \overline{B}]$ to maximize (20) taking $C_i, p, q, 1+r = p/q$ as given. This is equivalent to the problem

$$\max_{B_i \in [0,\overline{B}]} \frac{B_i}{q} - (1+r)B_i = \frac{1-p}{q} B_i$$

which has solution $B_i = \overline{B}$ if $p < 1$, $B_i = 0$ if $p > 1$, $B_i \in [0, \overline{B}]$ if $p = 1$. Note that this implies $(1-p)B_i = ((1-p)\overline{B})^+$. In the full game, i chooses C_i, B_i to maximize (20) given $p, q, 1+r = p/q$. The first order condition for C_i yields the Euler equation

$$\frac{\theta}{C_i} = (1+r)\left(\frac{B_i + \delta}{q} + (1+r)[p - C_i + B - B_i]\right)^{-1}$$

as shown in the main text. Rearranging and using $1 + r = p/q$,

$$\begin{aligned} C_i &= \frac{\theta}{1+\theta}\left(\frac{\delta}{p} + \frac{1-p}{p} B_i + [p + B]\right) \\ &= \frac{\theta}{1+\theta}\left(p + B + \frac{\delta}{p}\right) + \frac{\theta}{1+\theta}\left(\frac{1-p}{p}\overline{B}\right)^+ \end{aligned}$$

where the second line ((11) in the main text) uses $(1-p)B_i = ((1-p)\overline{B})^+$.

If $p = C < 1$ in equilibrium, then $B = \overline{B}$ and (11) becomes

$$\begin{aligned} C_i &= \frac{\theta}{1+\theta}\left(p + \overline{B} + \frac{\delta}{p}\right) + \frac{\theta}{1+\theta}\frac{1-p}{p}\overline{B} \\ C = \int C_i di &= \frac{\theta}{1+\theta}\left(C + \frac{\delta}{C}\right) + \frac{\theta}{1+\theta}\frac{1}{C}\overline{B} \\ C^2 &= \theta(\delta + \overline{B}) \end{aligned}$$

so $C = \sqrt{\theta(\overline{B}+\delta)} > 1$, a contradiction. If $p = C > 1$, then $B = 0$, and (11) becomes

$$\begin{aligned} C_i &= \frac{\theta}{1+\theta}\left(p + \frac{\delta}{p}\right) \\ C = \int C_i di &= \frac{\theta}{1+\theta}\left(C + \frac{\delta}{C}\right) \\ C^2 &= \theta\delta \end{aligned}$$

and $C = \sqrt{\theta\delta} < 1$, a contradiction. The only remaining possibility is $C = 1$. Substituting in (11),

$$1 = C = \int C_i di = \frac{\theta}{1+\theta}(1 + B + \delta)$$

implying $B = \theta^{-1} - \delta$ as claimed in the main text. □

A.9 Pricing game with finite number of farmers

First we show that in the continuum game, following any price profile $\mathbf{p} := \{p(j)\}_{j\in[0,1]}$, for any y, there exists a subgame equilibrium in which aggregate output is y and i's consumption is $c(i) = (p(i)/p)y(i) = (p(i)/p)^{1-\gamma}y$. Clearly this satisfies i's budget constraint and therefore is optimal. Integrating across i,

$$y = \int c(i)di = \int \left(\frac{p(i)}{p}\right)^{1-\gamma} diy$$

which is satisfied by definition of p. So we are done.

In the game with n farmers, consider the selection

$$c_i(\mathbf{p}) = p_{-i}^{-\gamma} p_i^{1-\gamma}\left(n^{-1}\sum_j p_j^{1-\gamma}\right)^{\frac{2\gamma-1}{1-\gamma}} y^*, \qquad y_i(\mathbf{p}) = p_{-i}^{1-\gamma} p_i^{-\gamma}\left(n^{-1}\sum_j p_j^{1-\gamma}\right)^{\frac{2\gamma-1}{1-\gamma}} y^*$$

where $1 = \frac{\widetilde{\gamma}}{\widetilde{\gamma}-1}(y^*)^\sigma v'(y^*)$, $\widetilde{\gamma} := \gamma - n^{-1}(2\gamma-1) < \gamma$. The derivatives of $c_i(\mathbf{p})$, $y_i(\mathbf{p})$ with respect

to p_i are

$$\frac{\partial c_i(\mathbf{p})}{\partial p_i} = \left[1 - \gamma + (2\gamma - 1)\frac{p_i^{1-\gamma}}{\sum_j p_j^{1-\gamma}}\right]\frac{c_i(\mathbf{p})}{p_i} \tag{21}$$

$$\frac{\partial y_i(\mathbf{p})}{\partial p_i} = \left[-\gamma + (2\gamma - 1)\frac{p_i^{1-\gamma}}{\sum_j p_j^{1-\gamma}}\right]\frac{y_i(\mathbf{p})}{p_i} \tag{22}$$

Farmer i's optimal pricing decision satisfies

$$c_i(\mathbf{p})^{-\sigma}\frac{\partial c_i(\mathbf{p})}{\partial p_i} = v'(y_i(\mathbf{p}))\frac{\partial y_i(\mathbf{p})}{\partial p_i}$$

$$c_i(\mathbf{p})^{-\sigma}\left[1 - \gamma + (2\gamma - 1)\frac{p_i^{1-\gamma}}{\sum_j p_j^{1-\gamma}}\right]\frac{c_i(\mathbf{p})}{p_i} = v'(y_i(\mathbf{p}))\left[-\gamma + (2\gamma - 1)\frac{p_i^{1-\gamma}}{\sum_j p_j^{1-\gamma}}\right]\frac{y_i(\mathbf{p})}{p_i}$$

$$\frac{p_i}{p_{-i}} = \frac{\gamma - (2\gamma - 1)\frac{p_i^{1-\gamma}}{\sum_j p_j^{1-\gamma}}}{\gamma - (2\gamma - 1)\frac{p_i^{1-\gamma}}{\sum_j p_j^{1-\gamma}} - 1}c_i(\mathbf{p})^{\sigma}v'(y_i(\mathbf{p})) \tag{23}$$

First, we argue that at any solution to farmer i's pricing problem, we must have $\frac{\partial c_i(\mathbf{p})}{\partial p_i} < 0$, $\frac{\partial y_i(\mathbf{p})}{\partial p_i} < 0$. Examining (21) and (22), we see that, given $\{p_j\}_{j\neq i}$, we have three cases depending on the position of p_i relative to cutoffs $\widetilde{p}_1 < \widetilde{p}_2$:

$$\frac{\partial c_i(\mathbf{p})}{\partial p_i} > 0, \frac{\partial y_i(\mathbf{p})}{\partial p_i} > 0 \text{ if } p_i < \widetilde{p}_1$$

$$\frac{\partial c_i(\mathbf{p})}{\partial p_i} > 0, \frac{\partial y_i(\mathbf{p})}{\partial p_i} < 0 \text{ if } \widetilde{p}_1 < p_i < \widetilde{p}_2$$

$$\frac{\partial c_i(\mathbf{p})}{\partial p_i} < 0, \frac{\partial y_i(\mathbf{p})}{\partial p_i} < 0 \text{ if } p_i > \widetilde{p}_2, \text{ where } \widetilde{p}_1 := \left(\frac{\gamma - 1}{\gamma\sum_{j\neq i} p_j^{1-\gamma}}\right)^{\frac{1}{\gamma-1}}, \widetilde{p}_2 := \left(\frac{\gamma}{(\gamma - 1)\sum_{j\neq i} p_j^{1-\gamma}}\right)^{\frac{1}{\gamma-1}}.$$

In other words, both c_i and y_i are first increasing and then increasing in p_i, but y_i peaks earlier. Also, note that $\lim_{p_i\to\infty} y_i(\mathbf{p}) = 0$. Take any price $p_i < \widetilde{p}_1$ and suppose it results in output and consumption (y_i, c_i). Since y_i is eventually decreasing in p_i, the same level of output is attainable with a higher price, say p_i', which implies a higher level of consumption, $(p_i'/p_i)c_i$. Since utility is increasing in consumption and decreasing in output, this contradicts the original allocation being optimal. So no $p_i < \widetilde{p}_1$ can be optimal. Clearly, no $p_i \in [\widetilde{p}_1, \widetilde{p}_2]$ can be optimal either, since a marginal increase in price would increase consumption and reduce output. So we must have $p_i > \widetilde{p}_2$ and $\frac{\partial c_i(\mathbf{p})}{\partial p_i} < 0$, $\frac{\partial y_i(\mathbf{p})}{\partial p_i} < 0$, as claimed.

Given this result, the numerator and denominator of the fraction on the right hand side of (23) (representing farmer i's markup) are both positive, so this markup is decreasing in p_i. Since the other terms on the right hand side are also increasing, while the left hand side is increasing, (23) uniquely defines i's optimal price. Further, note that this expression is homogeneous of degree zero in $\mathbf{p}$: for any $\lambda > 0$, if p_i is optimal given $\{p_j\}_{j\neq i}$, λp_i is optimal given $\{\lambda p_j\}_{j\neq i}$. Also, since

$\sum_j p_j^{1-\gamma}$ can be written as $p_i^{1-\gamma} + (n-1)p_{-i}^{1-\gamma}$, (23) implicitly defines p_i as a function of p_{-i}. The unique optimal solution to i's pricing problem is $p_i = p_{-i}$, since when evaluated at this value (23) reduces to

$$1 = \frac{\gamma - n^{-1}(2\gamma - 1)}{\gamma - n^{-1}(2\gamma - 1) - 1}(y^*)^\sigma v'(y^*)$$

which holds by definition of y^*. Finally, this implies that any Nash equilibrium features uniform pricing, $p_i = p$ for all i. For if by contradiction $p_j > p_i$, then

$$p_{-j} = \left(\varepsilon p_i^{1-\gamma} + \sum_{k \neq j} \varepsilon p_k^{1-\gamma}\right)^{\frac{1}{1-\gamma}} < \left(\varepsilon p_j^{1-\gamma} + \sum_{k \neq j} \varepsilon p_k^{1-\gamma}\right)^{\frac{1}{1-\gamma}} = p_{-i},$$

which contradicts the assumption that $p_j > p_i$ since $p_i = p_{-i}$, $p_j = p_{-j}$. So we are done.

Now consider the alternative selection

$$c_i(\mathbf{p})) = p_{-i}^{-\gamma} p_i^{1-\gamma} \left(n^{-1} \sum_j p_j^{1-\gamma}\right)^{\frac{2\gamma-1}{1-\gamma}} \widetilde{y} \prod_j p_j^{\chi}, \qquad y_i(\mathbf{p})) = p_{-i}^{1-\gamma} p_i^{-\gamma} \left(n^{-1} \sum_j p_j^{1-\gamma}\right)^{\frac{2\gamma-1}{1-\gamma}} \widetilde{y} \prod_j p_j^{\chi}$$

where $1 = \frac{\widetilde{\gamma} - \chi}{\widetilde{\gamma} - \chi - 1}(\widetilde{y})^\sigma v'(\widetilde{y})$. The optimal pricing decision now satisfies

$$\frac{p_i}{p_{-i}} = \frac{\gamma - \chi - (2\gamma - 1)\frac{p_i^{1-\gamma}}{\sum_j p_j^{1-\gamma}}}{\gamma - \chi - (2\gamma - 1)\frac{p_i^{1-\gamma}}{\sum_j p_j^{1-\gamma}} - 1} c_i(\mathbf{p})^\sigma v'(y_i(\mathbf{p})) \tag{24}$$

The same arguments as above imply that $\frac{\partial c_i(\mathbf{p})}{\partial p_i} < 0$, $\frac{\partial y_i(\mathbf{p})}{\partial p_i} < 0$ at any optimal price, so again, the right hand side of (24) is decreasing and this defines a unique optimal price p_i. When $p_j = 1$ for all $j \neq i$, this unique optimum is $p_i = 1$, since by definition of χ,

$$1 = \frac{\gamma - \chi - n^{-1}(2\gamma - 1)}{\gamma - \chi - n^{-1}(2\gamma - 1) - 1}(y^*)^\sigma v'(y^*)$$

So we are done. □

A.10 IOR game with finite number of traders

Following a history of goods market bids $(C_1, 1, ..., 1)$, payoffs can be rewritten

$$U_1(\mathbf{C}, \mathbf{B}) = \ln\left(\frac{C_1}{\varepsilon C_1 + 1 - \varepsilon}\right) + \ln\left(\frac{(1 - C_1)[\varepsilon(B_1 + \delta) + (1 - \varepsilon)(\varepsilon C_1 + 1 - \varepsilon)]}{\varepsilon B_1 + B_{-1} + \delta} + (\varepsilon C_1 + 1 - \varepsilon)\right)$$

$$U_i(\mathbf{C}, \mathbf{B}) = \ln\left(\frac{1}{\varepsilon C_1 + 1 - \varepsilon}\right) + \ln\left(\frac{\varepsilon(1 - C_1)[B_i - (\varepsilon C_1 + 1 - \varepsilon) + \delta]}{\varepsilon B_i + B_{-i} + \delta} + (\varepsilon C_1 + 1 - \varepsilon)\right) \text{ for } i > 1$$

The derivative of agent $i > 1$'s objective function with respect to B_i is proportional to

$$(1 - C_1)\frac{B_{-i} + (1-\varepsilon)\delta + \varepsilon(\varepsilon C_1 + 1 - \varepsilon)}{\varepsilon B_i + B_{-i} + \delta}$$

If $C_1 < 1$, this is positive for any feasible values of B_i, B_{-i}, and so $B_i = \overline{B}$ is optimal. If $C_1 > 1$, the term is negative, and $B_i = 0$ is optimal. Thus, if $C_1 < 1$, $B_{-1} = (1-\varepsilon)\overline{B}$, and trader 1's payoff becomes

$$\ln\left(\frac{C_1}{\varepsilon C_1 + 1 - \varepsilon}\right) + \ln\left((1 - C_1)\frac{[\varepsilon B_1 + (1-\varepsilon)(\varepsilon C_1 + 1 - \varepsilon)] + \varepsilon\delta}{\varepsilon B_1 + (1-\varepsilon)\overline{B} + \delta} + (\varepsilon C_1 + 1 - \varepsilon)\right)$$

Since $\varepsilon C_1 + 1 - \varepsilon < 1 < \overline{B}$, this is increasing in B_1, and trader 1 will also choose $B_1 = \overline{B}$.

If $C_1 > 1$, $B_{-1} = 0$, and trader 1's payoff is

$$\ln\left(\frac{C_1}{\varepsilon C_1 + 1 - \varepsilon}\right) + \ln\left((1 - C_1)\frac{[\varepsilon B_1 + (1-\varepsilon)(\varepsilon C_1 + 1 - \varepsilon)] + \varepsilon\delta}{\varepsilon B_1 + \delta} + (\varepsilon C_1 + 1 - \varepsilon)\right)$$

Since $\delta < 1$, this is increasing in B_1 for any ε. So again, trader 1 chooses $B_1 = \overline{B}$.

Having solved for $\mathbf{B}(\mathbf{C})$ following histories $\mathbf{C} = (C_1, 1, ..., 1)$, we substitute this selection into trader 1's objective function to characterize her optimal behavior given that all others set $C_i = 1$:

$$U_1(\mathbf{C}, \mathbf{B}(\mathbf{C})) = \begin{cases} \ln\left(\frac{C_1}{\varepsilon C_1 + 1 - \varepsilon}\right) + \ln\left(\frac{1-\varepsilon}{\overline{B}+\delta}(1 - C_1)(\varepsilon C_1 + 1 - \varepsilon) + 1\right) & \text{if } C_1 < 1 \\ 0 & \text{if } C_1 = 1 \\ \ln\left(\frac{C_1}{\varepsilon C_1 + 1 - \varepsilon}\right) + \ln\left(\frac{1-\varepsilon}{\varepsilon\overline{B}+\delta}(1 - C_1)(\varepsilon\overline{B} + \varepsilon C_1 + 1 - \varepsilon) + 1\right) & \text{if } C_1 > 1 \end{cases}$$

This function is continuous but not differentiable at $C_1 = 1$. To show that this bid is optimal, it suffices to show that the function is increasing for $C_1 < 1$, and decreasing for $C_1 > 1$. When $C_1 < 1$, its derivative is proportional to

$$\begin{aligned} &\frac{1}{C_1} - \frac{\varepsilon}{\varepsilon C_1 + 1 - \varepsilon} + \left(\frac{1-\varepsilon}{\overline{B}+\delta}(1 - C_1)(\varepsilon C_1 + 1 - \varepsilon) + 1\right)^{-1} \frac{1-\varepsilon}{\overline{B}+\delta}(-1 + 2\varepsilon(1 - C_1)) \\ =&(1-\varepsilon)\left[\frac{1}{C_1(\varepsilon C_1 + 1 - \varepsilon)} - \frac{1 - 2\varepsilon(1 - C_1)}{(1-\varepsilon)(1 - C_1)(\varepsilon C_1 + 1 - \varepsilon) + \overline{B} + \delta}\right] \end{aligned}$$

This expression is positive, as required: the first term in square brackets is greater than 1 while the second is less than 1 in absolute value. When $C_1 > 1$, the derivative is proportional to

$$\begin{aligned} &\frac{1}{C_1} - \frac{\varepsilon}{\varepsilon C_1 + 1 - \varepsilon} + \left(\frac{1-\varepsilon}{\varepsilon\overline{B}+\delta}(1 - C_1)(\varepsilon\overline{B} + \varepsilon C_1 + 1 - \varepsilon) + 1\right)^{-1} \frac{1-\varepsilon}{\varepsilon\overline{B}+\delta}\left(-\varepsilon\overline{B} - 1 + 2\varepsilon - 2\varepsilon C_1)\right) \\ =&(1-\varepsilon)\left[\frac{1}{C_1(\varepsilon C_1 + 1 - \varepsilon)} - \frac{\varepsilon\overline{B} + 1 + 2\varepsilon(C_1 - 1)}{\varepsilon\overline{B}[1 + (1-\varepsilon)(1 - C_1)] - (1-\varepsilon)(C_1 - 1)(\varepsilon C_1 + 1 - \varepsilon) + \delta}\right] \end{aligned}$$

The first term in square brackets is less than 1, while the numerator of the second fraction inside the

square brackets is greater than the denominator. (Note that the denominator cannot be negative at any optimal C_1, since the trader will never choose negative consumption.) Thus, the derivative is negative, as desired, and $C_1 = 1$ is indeed optimal. By symmetry, it follows that C_i is optimal for any agent i given that other traders play $C_j = 1$, for any selection $\mathbf{B}(\mathbf{C}^*)$.

A.11 Extensive form rationalizability in the pricing game

Here I define and characterize the set of strategies in the pricing game which are *extensive form rationalizable* (EFR) in the sense of Pearce (1984), drawing on the characterization in Battigalli (1997). First, some definitions. The *histories* or information sets at which each farmer i may have to act are $H := \{\varnothing\} \cup \mathcal{P}$, where $\varnothing$ denotes the initial history at the start of the morning subgame, and $\mathcal{P} := \mathbb{R}_+^{[0,1]}$ is the set of possible price profiles $\mathbf{p} := \{p_j\}_{j\in[0,1]}$, i.e. possible histories at the beginning of the afternoon subgame. A *strategy* s_i for farmer i is a pair $(p_i, c_i(\cdot))$ where $p_i \in \mathbb{R}_+$ and c_i is a function mapping $\mathcal{P}$ to $\mathbb{R}_+$; it describes what price p_i i sets in the morning, and her afternoon consumption $c_i(\mathbf{p})$ given any profile of prices. Let S^i be i's strategy set and $S = \prod_{i\in[0,1]} S^i$ the set of strategy profiles. We say a strategy $s_i = (p_i, c_i(\cdot))$ 'reaches' history $h = \{\hat{p}_j\}$ if $p_i = \hat{p}_i$ (that is, s_i is consistent with h realizing, if other agents act appropriately). A profile $s_{-i} = \{s_j\}_{j\neq i}$ reaches h if $p_j = \hat{p}_j$ for all $j \neq i$ (i.e. s_{-i} is consistent with h realizing if i acts appropriately). Every strategy and every profile of strategies reaches the initial history $h = \{\varnothing\}$.

We want to know what beliefs or conjectures it is 'reasonable' for i to entertain about others' behavior. In a dynamic setting, whether i is forming beliefs in a reasonable way depends on how she revises these beliefs upon observing new evidence: it might be reasonable ex ante to expect all other farmers to set $p_j = 1$, but it is absurd to maintain this belief after observing that they set $p_j = 2$. More precisely then, we want to know what 'belief functions' or updating systems are reasonable. In Battigalli (1997)'s terminology, a *consistent updating system* for i is a function $b^{-i}(h) = \{b^j(h)\}_{j\neq i} = \{\widetilde{p}^j(h), \widetilde{c}^j(\cdot|h)\}_{j\neq i}$ mapping histories $h \in H$ to (point) expectations about all other agents' strategies $s_{-i} := \{s_j\}_{j\neq i}$.[49] This must satisfy $\widetilde{p}^j(\mathbf{p}) = p_j$ for all $\mathbf{p}, j$: after observing a price profile, agents have correct expectations about prices. We also requre that for $h = \widetilde{\mathbf{p}}(\varnothing)$, $b^{-i}(h) = b^{-i}(\varnothing)$. That is, if i's initial expectation about the pricing profile is not falsified, she does not change her beliefs about any variables.

We define the set of rationalizable strategies inductively, as the outcome of an iterative process of reflection (Pearce, 1984). Farmer i is rational, and will only play strategies that are best responses to some possible strategy profile of others, 'deleting' those that are not. But i believes other farmers are rational, and will also delete such strategies; so i will only play best responses to strategy profiles that others may actually play, taking this into account. But i believes other farmers believe others are rational...and so on. Formally, let $R(0) = \{R^i(0)\}_{i\in[0,1]} = S$ and define $R(k)$ inductively: $s_i \in R^i(k)$ if $s_i \in R^i(k-1)$ and there exists a consistent updating system $b^{-i}(\cdot)$ such that

(i) for all $h \in H$, if s_i and $R^{-i}(k-1)$ reach h then $b^{-i}(h) \in R^{-i}(k-1)$;

[49] To reduce notation, we assume agents have point expectations about others' strategies. If we allowed agents to expect a distribution over others' strategies, as in Pearce (1984) and Battigalli (1997), the set of rationalizable strategies would be weakly larger – which can only strengthen our point that 'effective demand failures' are rationalizable.

(ii) s_i is a best response to $b^{-i}(h)$ given that history h has realized.

Condition (i) says that agents restrict their conjectures about others' behavior to strategies which have not been eliminated at an earlier stage, provided that some such strategy profile could have led to the observed history h. Condition (ii) says that strategies in $R^i(n)$ must be optimal given 'reasonable' beliefs. The set of rationalizable strategies is defined as $R(\infty) = \bigcap_{k\geq 0} R(k)$.

Having defined EFR, we now show that it does not rule out effective demand failures in which $y \neq y^*$. We claim that for any $(p_i, y) \in \mathbb{R}^2_+$, the strategy $(p_i, (p_i/p)^{1-\gamma}y)$ is rationalizable. Intuitively, these strategies are optimal if farmers have fixed expectations about afternoon output y, which they never revise, whatever the history of prices $\mathbf{p}$. Given y, call the set of such 'fixed-y' strategies $\widetilde{S}^i(y)$ and let $\widetilde{S}(y) = \prod_{i\in[0,1]} \widetilde{S}^i(y)$. Fix y and suppose $\widetilde{S}(y) \subset R(k-1)$; we will show $\widetilde{S}(y) \subset R(k)$, i.e. our reduction procedure never eliminates these strategies and $\widetilde{S}(y) \subset R(\infty)$.

Recall that i's optimal price and consumption, given aggregate variables p, y, are $p_i = pf(y)$ and $c_i = (p_j/p)^{1-\gamma}y$, respectively, for some increasing function f. Take any $s_i \in \widetilde{S}(y)$: we will rationalize this using the following updating system b^{-i}. At the initial history $h = \varnothing$, conjecture that farmer j's strategy is $\widetilde{p}_j = p_i/f(y)$ and $\widetilde{c}^j(\mathbf{p}) = (p_j/p)^{1-\gamma}y$ for each $j \neq i$. (In words, i expects that all other farmers will set prices equal to $\widetilde{p}_j$, and expects them to consume optimally, whatever the realized profile of prices.) Following any history $\mathbf{p}$, keep conjecturing $\widetilde{c}^j(\mathbf{p}) = (p_j/p)^{1-\gamma}y$, but update beliefs about other farmers' choice of p_j to be consistent with observed play.

This updating system satisfies condition (i): it conjectures a strategy for each $j \neq i$ that has not been eliminated by step $k-1$ of the reduction procedure. For it conjectures that each j playes a fixed-y strategy, and by assumption, these strategies have not yet been eliminated, $\widetilde{S}(y) \subset R(k-1)$. So we just need to check condition (ii), i.e. that $(p_i, (p_i/p)^{1-\gamma}y)$ is a best response to i's beliefs. After observing $\mathbf{p}$, i expects each farmer j to consume $(p_j/p)^{1-\gamma}y$; this implies that aggregate output will be $\int (p_j/p)^{1-\gamma}y dj = y$. i's consumption strategy $(p_i/p)^{1-\gamma}y$ is indeed a best response to this belief. Before observing $\mathbf{p}$, i still expects other farmers' consumption decisions to produce aggregate output y, but expects them to set prices $\widetilde{p}_j$. i's optimal response is $\widetilde{p}_j f(y) = p_i$, by definition of $\widetilde{p}_j$. So s_i is a best response to b^{-i}, and (ii) is satisfied; we are done. □

A.12 Learning in simple models

Given expectations x^e_t, y^e_t at the beginning of the date t 'morning', the realized aggregate action is $x_t = \alpha_{x,x}x^e_t + \alpha_{x,y}y^e_t$ and so the action taken in the afternoon is

$$\begin{aligned} y_t &= \alpha_{y,x}x_t + y^e_t \\ &= \alpha_{y,x}\alpha_{x,x}x^e_t + (1 + \alpha_{y,x}\alpha_{x,y})y^e_t \end{aligned}$$

Substituting into the learning equations, we can write the system as

$$\begin{bmatrix} x^e_{t+1} \\ y^e_{t+1} \end{bmatrix} = \begin{bmatrix} x^e_t \\ y^e_t \end{bmatrix} + g_t \left(\underbrace{\begin{bmatrix} \alpha_{x,x} & \alpha_{x,y} \\ \alpha_{y,x}\alpha_{x,x} & 1+\alpha_{y,x}\alpha_{x,y} \end{bmatrix}}_{A} \begin{bmatrix} x^e_t \\ y^e_t \end{bmatrix} - \begin{bmatrix} x^e_t \\ y^e_t \end{bmatrix} \right)$$

Theorems 7.1 and 7.2 in Evans and Honkapohja (2001) imply that this system is locally stable if all eigenvalues of A have real part less than 1, and locally unstable if A has an eigenvalue with real part bigger than 1. Thus for stability, we need

$$\begin{aligned} |A - I| &= (\alpha_{x,x} - 1)\alpha_{y,x}\alpha_{x,y} - \alpha_{x,y}\alpha_{y,x}\alpha_{x,x} > 0 \\ &= -\alpha_{y,x}\alpha_{x,y} > 0 \\ tr(A - I) &= \alpha_{x,x} - 1 + \alpha_{y,x}\alpha_{x,y} < 0 \end{aligned}$$

We can express this as $\alpha_{y,x}\alpha_{x,y} < \min\{0, 1 - \alpha_{x,x}\}$, as stated in the main text.

If $\alpha_{x,x} = 1$, $\alpha_{x,y} \neq 0$, and $\alpha_{y,x} = 0$, then $y^e_{t+1} = y_t = y^e_t$. For any initial conditions with $y^e_0 \neq 0$, y_t remains at that point forever and does not converge to 0.

A.13 Learning in the pricing game

Taking the limit $T \to \infty$, in the morning of date 0 farmer i solves

$$\begin{aligned} \max_{\{p_t(i), c_t(i)\}_{t=0}^{\infty}} & \sum_{t=0}^{\infty} \beta^t \left[\frac{c_t(i)^{1-\sigma}}{1-\sigma} - v\left(\left(\frac{p_t(i)}{p_t} \right)^{-\gamma} c_t \right) \right] \\ \text{s.t} & \sum_{t=0}^{\infty} p_t c_t(i) = \sum_{t=0}^{\infty} p_t(i) \left(\frac{p_t(i)}{p_t} \right)^{-\gamma} c_t, \end{aligned}$$

where p_t, c_t denote i's expectations about aggregate prices and consumption. In all the cases we consider, i will conjecture a constant rate of inflation Π and level of consumption c. Using this fact and defining $x_t(i) = p_t(i)/p_t$, we can rewrite the problem as

$$\begin{aligned} \max_{\{x_t(i), c_t(i)\}_{t=0}^{\infty}} & \sum_{t=0}^{\infty} \beta^t \left[\frac{c_t(i)^{1-\sigma}}{1-\sigma} - v\left(x_t(i)^{-\gamma} c \right) \right] \\ \text{s.t} & \sum_{t=0}^{\infty} \Pi^t c_t(i) = \sum_{t=0}^{\infty} \Pi^t x_t(i)^{1-\gamma} c \end{aligned}$$

Attaching a multiplier λ to the constraint, the first order optimality conditions are

$$\beta^t c_t(i)^{-\sigma} = \lambda \Pi^t \tag{25}$$

$$x_t(i) = \frac{\gamma}{\gamma - 1} \left(\frac{\beta}{\Pi} \right)^t \frac{v'(x_t(i)^{-\gamma} c)}{\lambda} \tag{26}$$

Approximating $v'(y) \approx v_0 y^{\varphi}$, using (26) to substitute out for $x_t(i)^{1-\gamma}$ in the budget constraint, and rearranging, we have

$$\lambda^{-\frac{1}{1+\gamma\varphi}} = \phi_0 c^{\frac{1+\varphi}{(1+\gamma\varphi)((1+\gamma\varphi)/\sigma+\gamma-1)}} \left(\frac{1-\Pi^{1-1/\sigma}\beta^{1/\sigma}}{1-\Pi\left(\frac{\beta}{\Pi}\right)^{\frac{1-\gamma}{1+\gamma\varphi}}} \right)^{\frac{1}{(1+\gamma\varphi)/\sigma+\gamma-1}}$$

where ϕ_0 is a constant. Substituting this into (26):

$$x_0(i) = \left(\frac{1-\Pi^{1-1/\sigma}\beta^{1/\sigma}}{1-\Pi\left(\frac{\beta}{\Pi}\right)^{\frac{1-\gamma}{1+\gamma\varphi}}} \right)^{\frac{\sigma}{1-\sigma+\gamma(\varphi+\sigma)}} \left(\frac{c}{y^*} \right)^{\frac{\sigma+\varphi}{1-\sigma+\gamma(\varphi+\sigma)}}$$

$$\frac{p_0(i)}{p_{-1}} = \Pi \left(\frac{1-\Pi^{1-1/\sigma}\beta^{1/\sigma}}{1-\Pi\left(\frac{\beta}{\Pi}\right)^{\frac{1-\gamma}{1+\gamma\varphi}}} \right)^{\frac{\sigma}{1-\sigma+\gamma(\varphi+\sigma)}} \left(\frac{c}{y^*} \right)^{\frac{\sigma+\varphi}{1-\sigma+\gamma(\varphi+\sigma)}}$$

Log-linearizing and aggregating across farmers, we have

$$\pi_0 = \left[1 + \frac{\beta}{1-\beta}\frac{1}{1+\gamma\varphi} \right] \pi_0^m + \frac{\sigma+\varphi}{1-\sigma+\gamma(\varphi+\sigma)} c_0^m$$

where π_0^m and c_0^m denote log-deviations of the representative farmer's conjectures regarding aggregate inflation and consumption in the morning of period 0. Since nothing is special about period 0, we can replace the subscript 0 with t, yielding (14) in the main text.

When choosing $c_0(i)$ in the afternoon of date 0, the farmer's problem is unchanged except that she does not choose $x_0(i)$, but instead takes $x_0(i) = 1$ as given. (We assume all farmers have the same expectations, so they always find out ex post that they set the same prices, even if they had intended to set their prices higher or lower than the aggregate.) Using (26) to substitute out for $x_t(i)$ in the budget constraint for $t \geq 1$ and rearranging,

$$\lambda^{-1/\sigma} \frac{1}{1-\Pi^{1-1/\sigma}\beta^{1/\sigma}} 1 = c + \phi_1 \lambda^{-\frac{1-\gamma}{1+\gamma\varphi}} c^{1+\frac{(1-\gamma)\varphi}{1+\gamma\varphi}} \frac{1}{\Pi^{-1}\left(\frac{\Pi}{\beta}\right)^{\frac{1-\gamma}{1+\gamma\varphi}} - 1}$$

where ϕ_1 is a constant. Log-linearizing and rearranging,

$$\widehat{c}_0(i) = -\frac{1}{\sigma}\widehat{\lambda} = \frac{1+\gamma\varphi-\beta(\gamma-1)\varphi}{1+\gamma\varphi+\beta(\gamma-1)\sigma}\widehat{c} + \frac{1}{1-\beta}\frac{1}{\sigma}\frac{(1-\sigma)(1+\gamma\varphi)+\beta\gamma\sigma(1+\varphi)}{1+\gamma\varphi+\beta(\gamma-1)\sigma}\pi$$

Aggregating across farmers, noting again that there is nothing special about date 0, and using π_t^a and c_t^a to denote farmers' expectations in the afternoon of period t, we have (15) in the main text.

We can rewrite the system (14)-(15) as

$$\pi_t = \alpha_{\pi,\pi}\pi_t^m + \alpha_{\pi,c}c_t^m$$
$$\widehat{c}_t = \alpha_{c,\pi}\pi_t^a + \alpha_{c,c}c_t^a$$

Note that all the α coefficients are positive and $\alpha_{\pi,\pi} > 1$. Using the learning rules (13), we can rewrite the system in terms of only the 'morning' expectations:

$$\pi_{t+1}^m = \pi_t^m + g_t(\alpha_{\pi,\pi}\pi_t^m + \alpha_{\pi,c}c_t^m - \pi_t^m)$$
$$c_{t+1}^m = c_t^m + g_t(\alpha_{c,\pi}\pi_t^m + \alpha_{c,c}c_t^m + \alpha_{c,\pi}g_t(\alpha_{\pi,\pi}\pi_t^m + \alpha_{\pi,c}c_t^m - \pi_t^m) - c_t^m)$$

To apply Theorem 7.2 in Evans and Honkapohja (2001), define $\theta = (\pi_t^m, c_t^m)$ and $M(\theta_t, g_t)$ by

$$M(\theta_t, g_t) = \begin{bmatrix} \alpha_{\pi,\pi}\pi_t^m + \alpha_{\pi,c}c_t^m \\ \alpha_{c,\pi}\pi_t^m + \alpha_{c,c}c_t^m + \alpha_{c,\pi}g_t(\alpha_{\pi,\pi}\pi_t^m + \alpha_{\pi,c}c_t^m - \pi_t^m) \end{bmatrix}$$

Clearly M is continuously differentiable and satisfies $M(0, g_t) = \mathbf{0}$. Its Jacobian at $(0,0)$ and $(0, g_t)$, respectively, is

$$D_1M(0,0) = \begin{bmatrix} \alpha_{\pi,\pi} & \alpha_{\pi,c} \\ \alpha_{c,\pi} & \alpha_{c,c} \end{bmatrix}$$
$$D_1M(0,g_t) = \begin{bmatrix} \alpha_{\pi,\pi} & \alpha_{\pi,c} \\ \alpha_{c,\pi}(1-(1-\alpha_{\pi,\pi})g_t) & \alpha_{c,c} + \alpha_{c,\pi}\alpha_{\pi,c}g_t \end{bmatrix}$$

We need to check that $D_1M(0,g_t)$ does not have an eigenvalue equal to $-(1-g_t)/g_t$. The determinant $\left|D_1M(0,g_t) + \frac{1-g_t}{g_t}\right|$ equals

$$\begin{vmatrix} \alpha_{\pi,\pi} + \frac{1-g_t}{g_t} & \alpha_{\pi,c} \\ \alpha_{c,\pi}(1-(1-\alpha_{\pi,\pi})g_t) & \alpha_{c,c} + \alpha_{c,\pi}\alpha_{\pi,c}g_t + \frac{1-g_t}{g_t} \end{vmatrix}$$
$$= \left(\alpha_{\pi,\pi} + \frac{1-g_t}{g_t}\right)\left(\alpha_{c,c} + \alpha_{c,\pi}\alpha_{\pi,c}g_t + \frac{1-g_t}{g_t}\right) - \alpha_{\pi,c}\alpha_{c,\pi}(1-(1-\alpha_{\pi,\pi})g_t)$$
$$= \alpha_{\pi,\pi}\alpha_{c,c} + \frac{1-g_t}{g_t}\alpha_{c,c} + \alpha_{\pi,\pi}\alpha_{c,\pi}\alpha_{\pi,c}g_t + \frac{1-g_t}{g_t}\alpha_{c,\pi}\alpha_{\pi,c}g_t + \alpha_{\pi,\pi}\frac{1-g_t}{g_t}$$
$$+ \frac{(1-g_t)^2}{g_t^2} - \alpha_{\pi,c}\alpha_{c,\pi} + \alpha_{\pi,c}\alpha_{c,\pi}(1-\alpha_{\pi,\pi})g_t$$
$$= \alpha_{\pi,\pi}\alpha_{c,c} + \frac{1-g_t}{g_t}\alpha_{c,c} + \alpha_{\pi,\pi}\frac{1-g_t}{g_t} + \frac{(1-g_t)^2}{g_t^2} > 0$$

since $\alpha_{\pi,\pi}, \alpha_{c,c} > 0$. So indeed, it does not have such an eigenvalue. By Theorem 7.2 in Evans and Honkapohja (2001), then, $(0,0)$ is locally unstable if $D_1M(0,0)$ has an eigenvalue with real part greater than 1. Let $p(\lambda) = |D_1M(0,0) - \lambda I|$ be the characteristic polynomial; we have $p(\alpha_{\pi,\pi}) = -\alpha_{\pi,c}\alpha_{c,\pi} < 0$ (recall all the αs are positive), but $p(\lambda) \to \infty$ as $\lambda \to \infty$. So the system

has a real eigenvalue $\lambda > \alpha_{\pi,\pi} > 1$, and we are done. □

A.14 PRE in noisy REE model

A rational trader who knows that s is contained in an interval $S = (s_1, s_2]$ seeks to maximize

$$V(x|p,S) = -\frac{1}{\alpha}\mathbb{E}\left[\exp\left(-\alpha[(\theta-p)x]\right) \mid s \in S\right] = -\frac{1}{\alpha}\mathbb{E}\left[\mathbb{E}[\exp\left(-\alpha[(\theta-p)x]\right)|s] \mid s \in S\right]$$

Defining $\sigma_s^2 = \sigma^2 + \sigma_\varepsilon^2$, we have

$$\theta \mid s \sim N\left(\frac{\sigma_\varepsilon^2\overline{\theta}+\sigma^2 s}{\sigma_s^2}, \frac{\sigma^2\sigma_\varepsilon^2}{\sigma_s^2}\right),\ -\alpha\theta x \mid s \sim N\left(-\alpha x\frac{\sigma_\varepsilon^2\overline{\theta}+\sigma^2 s}{\sigma_s^2}, \frac{\alpha^2x^2\sigma^2\sigma_\varepsilon^2}{\sigma_s^2}\right)$$

Thus, we have

$$\mathbb{E}[\exp\left(-\alpha\theta x\right)|s] = \exp\left(-\alpha x\frac{\sigma_\varepsilon^2\overline{\theta}+\sigma^2 s}{\sigma_s^2} + \frac{1}{2}\frac{\alpha^2x^2\sigma^2\sigma_\varepsilon^2}{\sigma_s^2}\right)$$

We can rewrite the trader's objective function as

$$V(x|p,S) = -\frac{1}{\alpha}\exp\left(-\alpha\left[\frac{\sigma_\varepsilon^2}{\sigma_s^2}\overline{\theta}-p\right]x + \frac{\alpha^2}{2}\frac{\sigma^2\sigma_\varepsilon^2}{\sigma_s^2}x^2\right)\mathbb{E}\left[\exp\left(-\frac{\alpha\sigma^2 s}{\sigma_s^2}x\right) \mid s \in S\right]$$

Define $\beta = \frac{\alpha\sigma^2 x}{\sigma^2}$. We have

$$\mathbb{E}\left[\exp(-\beta s)|s \in S\right] = \frac{1}{\Phi\left(\frac{s_2-\overline{\theta}}{\sigma_s}\right) - \Phi\left(\frac{s_1-\overline{\theta}}{\sigma_s}\right)}\int_{s_1}^{s_2}\exp(-\beta s)\frac{1}{\sigma_s\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{s-\overline{\theta}}{\sigma_s}\right)^2\right)ds$$

Ignoring the normalizing constant (which does not depend on x and will prove irrelevant to the trader's maximization problem), we can complete the square to rewrite the integral as

$$\begin{aligned}
&\int_{s_1}^{s_2}\exp(-\beta s)\frac{1}{\sigma_s\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{s-\overline{\theta}}{\sigma_s}\right)^2\right)ds \\
=&\int_{s_1}^{s_2}\frac{1}{\sigma_s\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{s^2-2[\overline{\theta}-\beta\sigma_s^2]s+\overline{\theta}^2}{\sigma_s^2}\right)\right)ds \\
=&\exp\left(-\overline{\theta}\beta+\frac{\beta^2\sigma_s^2}{2}\right)\int_{s_1}^{s_2}\frac{1}{\sigma_s\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{s-[\overline{\theta}-\beta\sigma_s^2]}{\sigma_s}\right)^2\right)ds \\
=&\exp\left(-\overline{\theta}\beta+\frac{\beta^2\sigma_s^2}{2}\right)\left[\Phi\left(\frac{s_2-(\overline{\theta}-\beta\sigma_s^2)}{\sigma_s}\right)-\Phi\left(\frac{s_1-(\overline{\theta}-\beta\sigma_s^2)}{\sigma_s}\right)\right]
\end{aligned}$$

Thus, ignoring the irrelevant normalizing constant and using $\sigma_s^2 = \sigma^2 + \sigma_\varepsilon^2$,

$$V(x|p,S) \propto -\frac{1}{\alpha}\exp\left(-\alpha(\overline{\theta}-p)x+\frac{\alpha^2}{2}\sigma^2x^2\right)\left[\Phi\left(\frac{s_2-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)-\Phi\left(\frac{s_1-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)\right]$$

When $S = (-\infty, +\infty)$, the last term drops out and this naturally reduces to the objective function in the non-revealing equilibrium.

In the PRE, when $S = (c, \infty)$, this becomes

$$-\frac{1}{\alpha}\exp\left(-\alpha(\overline{\theta}-p)x+\frac{\alpha^2}{2}\sigma^2x^2\right)\left[1-\Phi\left(\frac{c-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)\right]$$

and the first order condition is

$$\overline{\theta}-p-\alpha\sigma^2x+\frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi\left(\frac{c-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)}{1-\Phi\left(\frac{c-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)}=0$$

Given prices, knowing that $s > c$ increases the demand for the risky asset, compared to the case where the rational trader has no information. When $S = (-\infty, c]$, the objective function is instead

$$-\frac{1}{\alpha}\exp\left(-\alpha(\overline{\theta}-p)x+\frac{\alpha^2}{2}\sigma^2x^2\right)\Phi\left(\frac{c-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)$$

and the optimality condition becomes

$$\overline{\theta}-p-\alpha\sigma^2x-\frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi\left(\frac{c-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)}{\Phi\left(\frac{c-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)}=0$$

Let c be a linear function of $x = y-u$, say $c(x) = c_0+c_1x$. We would like to construct decreasing functions $\underline{p}(x)$, $\overline{p}(x)$ such that $\lim_{x\to-\infty}\underline{p}(x) = \overline{\theta}$, $\lim_{x\to+\infty}\underline{p}(x) = -\infty$, $\lim_{x\to-\infty}\overline{p}(x) = +\infty$, $\lim_{x\to+\infty}\overline{p}(x) = \overline{\theta}$.

Define $\Psi_+(a) = \mathbb{E}[\theta|\theta > a] = \sigma\frac{\phi(a/\sigma)}{1-\Phi(a/\sigma)}$. Its derivative $\Psi'_+(a)$ is positive and bounded on $(0,1)$, and $\lim_{a\to-\infty}\Psi_+(a) = 0$, $\lim_{a\to+\infty}\Psi_+(a) = +\infty$. If we define $\Psi_-(a) = \mathbb{E}[\theta|\theta \le a] = -\sigma\frac{\phi(a/\sigma)}{\Phi(a/\sigma)}$, then since $\Psi_-(a) = -\Psi_+(-a)$, it follows that $\Psi'_-(a)$ is positive and bounded on $(0,1)$ and $\lim_{a\to-\infty}\Psi_-(a) = -\infty$, $\lim_{a\to+\infty}\Psi_-(a) = 0$.

Using these definitions, the optimality conditions can be written

$$\underline{p}(x) = \overline{\theta}-\alpha\sigma^2x+\frac{\sigma^2}{\sigma_s^2}\Psi_-(c_0+c_1x-\overline{\theta}+\alpha\sigma^2x)$$
$$\overline{p}(x) = \overline{\theta}-\alpha\sigma^2x+\frac{\sigma^2}{\sigma_s^2}\Psi_+(c_0+c_1x-\overline{\theta}+\alpha\sigma^2x)$$

Taking derivatives,

$$\underline{p}'(x) = -\alpha\sigma^2+\frac{\sigma^2}{\sigma_s^2}(c_1+\alpha\sigma^2)\Psi'_-(c_0+c_1x-\overline{\theta}+\alpha\sigma^2x)$$
$$\overline{p}'(x) = -\alpha\sigma^2+\frac{\sigma^2}{\sigma_s^2}(c_1+\alpha\sigma^2)\Psi'_+(c_0+c_1x-\overline{\theta}+\alpha\sigma^2x)$$

Since Ψ'_-, Ψ'_+ are bounded in $(0,1)$, both functions will be decreasing provided that $c_1 \leq \alpha\sigma_\varepsilon^2$. It remains to check the limits. We can rewrite

$$\underline{p}(x) = \overline{\theta} - \frac{\sigma^2}{\sigma_s^2} \frac{\alpha\sigma_s^2 x \Phi\left(\frac{c_0+c_1x-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right) + \sigma_s\phi\left(\frac{c_0+c_1x-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)}{\Phi\left(\frac{c_0+c_1x-\overline{\theta}+\alpha\sigma^2x}{\sigma_s}\right)}$$

As $x \to -\infty$, if $c_1 + \alpha\sigma^2 \leq 0$, $\underline{p}(x) \to +\infty$ which is not what we want. Assume then that $c_1 + \alpha\sigma^2 > 0$; then both numerator and denominator converge to zero. Applying L'Hôpital's rule, the derivative of the numerator is

$$\begin{aligned}
&\alpha\sigma_s^2\Phi\left(\frac{c_0 + c_1x - \overline{\theta} + \alpha\sigma^2x}{\sigma_s}\right) \\
&+\alpha\sigma_s x(c_1+\alpha\sigma^2)\phi\left(\frac{c_0 + c_1x - \overline{\theta} + \alpha\sigma^2x}{\sigma_s}\right) + (c_1+\alpha\sigma^2)\phi'\left(\frac{c_0 + c_1x - \overline{\theta} + \alpha\sigma^2x}{\sigma_s}\right)
\end{aligned}$$

Assuming $c_0 = \overline{\theta}$, this becomes

$$\begin{aligned}
&\alpha\sigma_s^2\Phi\left(\frac{c_1x + \alpha\sigma^2x}{\sigma_s}\right) + \alpha\sigma_s x(c_1+\alpha\sigma^2)\phi\left(\frac{c_1x + \alpha\sigma^2x}{\sigma_s}\right) + (c_1+\alpha\sigma^2)\phi'\left(\frac{c_1x + \alpha\sigma^2x}{\sigma_s}\right) \\
&= \alpha\sigma_s^2\Phi\left(\frac{c_1x + \alpha\sigma^2x}{\sigma_s}\right) + \alpha\sigma_s x(c_1+\alpha\sigma^2)\phi\left(\frac{c_1x + \alpha\sigma^2x}{\sigma_s}\right) - \frac{(c_1+\alpha\sigma^2)^2x}{\sigma_s}\phi\left(\frac{c_1x + \alpha\sigma^2x}{\sigma_s}\right)
\end{aligned}$$

The last two terms will cancel if

$$\begin{aligned}
\alpha\sigma_s(c_1+\alpha\sigma^2) - \frac{(c_1+\alpha\sigma^2)^2}{\sigma_s} &= 0 \\
\alpha\sigma_s^2 - (c_1+\alpha\sigma^2) &= 0 \\
c_1 &= \alpha\sigma_\varepsilon^2
\end{aligned}$$

Imposing $c_1 = \alpha\sigma_\varepsilon^2$ (which satisfies $c_1+\alpha\sigma^2 > 0$ as assumed above), the numerator and denominator have derivatives $\alpha\sigma_s^2\Phi(\alpha\sigma_s x)$ and $\alpha\sigma_s\phi(\alpha\sigma_s x)$ respectively. Both these terms converge to zero as $x \to -\infty$. Applying L'Hôpital again, the second derivatives of the numerator and denominator are $\alpha^2\sigma_s^3\phi(\alpha\sigma_s x)$ and $\alpha^2\sigma_s^2\phi'(\alpha\sigma_s x) = -\alpha^3\sigma_s^3 x\phi(\alpha\sigma_s x)$ respectively. Their ratio is

$$\frac{\alpha^2\sigma_s^3\phi(\alpha\sigma_s x)}{-\alpha^3\sigma_s^3 x\phi(\alpha\sigma_s x)} = -\frac{1}{\alpha x} \to 0 \text{ as } x \to -\infty.$$

So, with the cutoff $c(x) = \overline{\theta} + \alpha\sigma_\varepsilon^2$, we indeed have $\lim_{x\to-\infty}\underline{p}(x) = \overline{\theta}$.

As $x \to +\infty$, $\Phi(\alpha\sigma_s x) \to 1$ and so we have $\underline{p}(x) \to -\infty$, as desired.

Similarly, we can write

$$\begin{aligned}\overline{p}(x) &= \overline{\theta} - \alpha\sigma^2 x + \frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi\left(\alpha\sigma_s x\right)}{1-\Phi\left(\alpha\sigma_s x\right)} \\ &= \overline{\theta} + \frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi\left(\alpha\sigma_s x\right) - \alpha\sigma_s^2 x[1-\Phi\left(\alpha\sigma_s x\right)]}{1-\Phi\left(\alpha\sigma_s x\right)} \\ &= \overline{\theta} + \frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi\left(-\alpha\sigma_s x\right) + \alpha\sigma_s^2(-x)\Phi\left(-\alpha\sigma_s x\right)}{\Phi\left(-\alpha\sigma_s x\right)}\end{aligned}$$

Note that $\overline{p}(x) - \overline{\theta} = -(\underline{p}(-x) - \theta)$. Thus, it follows that $\lim_{x\to-\infty}\overline{p}(x) = +\infty$, $\lim_{x\to+\infty}\overline{p}(x) = \overline{\theta}$, as desired. So we are done: as claimed in the main text, the equilibrium price function is

$$p(s,u) = \begin{cases}\overline{\theta} - \alpha\sigma^2(y-u) - \frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi(\alpha\sigma_s(y-u))}{\Phi(\alpha\sigma_s(y-u))} & \text{if } s \leq \overline{\theta} + \alpha\sigma_\varepsilon^2(y-u) := s(u) \\ \overline{\theta} - \alpha\sigma^2(y-u) + \frac{\sigma^2}{\sigma_s^2}\frac{\sigma_s\phi(\alpha\sigma_s(y-u))}{1-\Phi(\alpha\sigma_s(y-u))} & \text{if } s > s(u).\end{cases}$$

□

A.15 Rational inattention

Theorem 1 in Matějka and McKay (2015) implies that $p(z_L), p(z_H)$ satisfy

$$p(z_j) = \frac{pe^{v(y(z_j))/\lambda}}{1-p+pe^{v(y(z_j))/\lambda}} = \frac{pa_j}{1-p+pa_j}$$

where p is the unconditional probability of choosing $x_i = 1$ and we define $a_j = e^{v(y(z_j))/\lambda}$ to economize on notation. Lemma 2 in Matějka and McKay (2015) implies that p maximizes

$$\frac{1}{2}\lambda\ln\left(1-p+pa_L\right) + \frac{1}{2}\lambda\ln\left(1-p+pa_H\right)$$

implying

$$\begin{aligned}p &= \frac{2-a_H-a_L}{2(a_H-1)(a_L-1)} \\ p(z_j) &= \frac{a_j}{\frac{2(a_H-1)(a_L-1)}{2-a_H-a_L}+a_j-1} = \frac{a_j}{\frac{2a_Ha_L-2}{2-a_H-a_L}+1+a_j} \\ &= \frac{\exp\left(\frac{v(y(z_j))}{\lambda}\right)}{\frac{2\exp\left(\frac{v(y(z_H))+v(y(z_L))}{\lambda}\right)-2}{2-\exp\left(\frac{v(y(z_H))}{\lambda}\right)-\exp\left(\frac{v(y(z_L))}{\lambda}\right)}+1+\exp\left(\frac{v(y(z_j))}{\lambda}\right)}\end{aligned}$$

In equilibrium, since $p(z_j) = z_j$, $p = \overline{z} = \frac{z_L + z_H}{2}$, we have

$$\begin{aligned} z_j &= \frac{\overline{z} a_j}{1 - \overline{z} + \overline{z} a_j} \\ \exp\left(\frac{v(y(z_j))}{\lambda}\right) = a_j &= \frac{z_j/(1-z_j)}{\overline{z}/(1-\overline{z})} \\ \left(\frac{1-y(z_j)}{y(z_j)}\right)^{1/\lambda} &= \frac{z_j/(1-z_j)}{\overline{z}/(1-\overline{z})} \\ y(z_j) &= \frac{\left(\frac{\overline{z}}{1-\overline{z}}\right)^{\lambda}}{\left(\frac{z_H}{1-z_H}\right)^{\lambda} + \left(\frac{\overline{z}}{1-\overline{z}}\right)^{\lambda}} \end{aligned}$$

as in the main text.

A.16 DE and CD

Assuming that date t investors perceive z_{t+1} to be normally distributed with mean $E_t^i[z_{t+1}]$ and variance $V_t^i[z_{t+1}]$, we can rewrite their objective function as

$$E_t^i\left[-\frac{1}{\alpha}\exp(-\alpha(z_{t+1} - q_t)k_t)\right] = -\frac{1}{\alpha}\exp\left(-\alpha(E_t^i[z_{t+1}] - q_t)k_t + \frac{\alpha^2}{2}V_t^i[z_{t+1}]k_t^2\right)$$

Maximizing this with respect to k_t yields the demand for capital described in the main text, $k_t = (\alpha V_t^i[z_{t+1}])^{-1}(E_t^i[z_{t+1}] - q_t)$; equating demand and supply yields the equilibrium price described there. The characterization of fully revealing equilibria under RE and CD follows immediately. It also follows that if there is a revealing equilibrium in DE, the price is

$$q_t = \frac{\rho(1+\theta)z_t - \theta\rho^2 z_{t-1}}{\alpha\sigma^2 + \nu} \tag{27}$$

To see that one can indeed infer z_t from the history $\{q_{t-\ell}\}_{\ell=0}^{\infty}$ using (17), substitute (27) into (17):

$$\begin{aligned} z_t &= \frac{\alpha\sigma^2 + \nu}{\rho(1+\theta)}\sum_{\ell=0}^{\infty}\left(\frac{\theta\rho}{1+\theta}\right)^{\ell}\frac{\rho(1+\theta)z_{t-\ell} - \theta\rho^2 z_{t-\ell-1}}{\alpha\sigma^2 + \nu} \\ &= \left[\sum_{\ell=0}^{\infty}\left(\frac{\theta\rho}{1+\theta}\right)^{\ell} z_{t-\ell} - \sum_{\ell=0}^{\infty}\left(\frac{\theta\rho}{1+\theta}\right)^{\ell}\frac{\theta\rho}{1+\theta}z_{t-\ell-1}\right] = z_t, \text{ as required.} \end{aligned}$$

A.17 Shallow reasoning and level-k

Shallow reasoning We define an equilibrium of the IOR game with shallow reasoning (Angeletos and Sastry, 2021) (or what Angeletos and Lian (2023) call heterogeneous priors) as a collection of functions $\{\{C_i(\theta), B_i(\theta, p)\}_{i\in[0,1]}, C(\theta), B(\theta)\}$. In particular, $B_i(\theta, p)$ denotes the bond market bid that informed trader i would place in state θ after observing the realized price p, which equals the aggregate goods market bid $C(\theta)$ – which may, in principle, differ from her prior expectation of this

aggregate bid, namely $\lambda C(\theta) + (1-\lambda)C(1)$. An equilibrium is a collection satisfying the following conditions:

1. for each trader i, realization θ and price level p, $B_i(\theta, p) \in [0, \overline{B}]$ maximizes $(1-p)B_i$.

2. for each i and θ, $C_i(\theta)$ maximizes

$$\theta \ln\left(\frac{C_i}{\widetilde{p}}\right) + \ln\left(\frac{B_i(\theta,\widetilde{p})}{\widetilde{q}} + (1+\widetilde{x})[\widetilde{p} - C_i + \widetilde{B} - B_i(\theta,\widetilde{p})]\right) \tag{28}$$

given expectations for prices and aggregates

$$\begin{aligned}
\widetilde{p} &= \lambda C(\theta) + (1-\lambda)C(1) \\
\widetilde{B} &= \lambda \int B_i(\theta,\widetilde{p})di + (1-\lambda)\int B_i(1,\widetilde{p})di \\
\widetilde{q} &= \widetilde{B} + \delta \\
1 + \widetilde{x} &= \widetilde{p}/\widetilde{q}
\end{aligned}$$

3. $C(\theta) = \int C_i(\theta)di$, $B(\theta) = \int B_i(\theta, C(\theta))di$.

When implementing shallow reasoning/heterogeneous priors in the extensive form IOR game, in principle we need to take a stand on two issues which do not arise in the static models studied by Angeletos and Sastry (2021); Angeletos and Lian (2023). First, how do informed traders update their beliefs in the afternoon after observing actions taken in the morning (which might reveal that all other agents are in fact informed)? We sidestep this question by noting that in this particular game, behavior in the afternoon subgame is nonstrategic: i's optimal choice of B_i depends on the aggregate bid $C = p$ placed in the morning subgame (which we assume informed agents observe), but not on other agents' behavior in this subgame. Thus, in condition 1 above, we do not take a stand on whether i updates her beliefs about the fraction of informed traders. Second, how do informed traders anticipate that *uninformed* traders will update their beliefs after observing the aggregate bid $C = p$ placed in the morning subgame? We assume they anticipate that uninformed traders behave in the same way as an informed trader would behave if $\theta = 1$.

Using our earlier characterization of optimizing behavior in this game,

$$C(\theta) = \int_i C_i(\theta)di = \frac{\theta}{1+\theta}\left(\widetilde{p} + \widetilde{B}\right) + \frac{\theta}{1+\theta}\left(\frac{1-\widetilde{p}}{\widetilde{p}}\overline{B}\right)^+$$

Since $C(1) = 1$, $\widetilde{p} = \lambda C(\theta) + (1-\lambda)C(1)$ will be greater than (less than) 1 if and only if $C(\theta)$ is greater than (less than) 1. Suppose $C(\theta) < 1$, $\widetilde{p} < 1$. Then $\widetilde{B} = \overline{B}$ and

$$C(\theta) = \frac{\theta}{1+\theta}\left(\lambda C(\theta) + 1 - \lambda + \overline{B}\right) + \frac{\theta}{1+\theta}\left(\frac{1}{\lambda C(\theta) + 1 - \lambda} - 1\right)\overline{B}$$

$$\left(C(\theta) - \theta(1-\lambda)(1-C(\theta))\right)\left(\lambda C(\theta) + 1 - \lambda\right) = \theta\overline{B} > 1,$$

given our assumptions on θ, which is a contradiction since the left hand side of this equation is less than 1. Suppose instead $C(\theta), \widetilde{p} > 1$. Then $\widetilde{B} = 0$ and

$$\begin{aligned} C(\theta) &= \frac{\theta}{1+\theta}\left(\lambda C(\theta) + 1 - \lambda\right) \\ &= \frac{\theta(1-\lambda)}{1+\theta(1-\lambda)} < 1, \end{aligned}$$

a contradiction. So we must have $C(\theta) = \widetilde{p} = 1$, implying

$$1 = \frac{\theta}{1+\theta}(1 + \widetilde{B})$$

i.e. $\widetilde{B} = \theta^{-1}$. When $\theta = 1$, this implies $\widetilde{B} = \int B_i(1,1)di = 1$. For $\theta \neq 1$, we must have

$$\begin{aligned} \widetilde{B} = \theta^{-1} &= \lambda \int B_i(\theta, 1)di + 1 - \lambda \\ B(\theta) := \int B_i(\theta,1)di &= \lambda^{-1}(\theta^{-1} - 1) + 1, \end{aligned}$$

as stated in the main text.

Level k Define $z = \frac{\theta}{1+\theta}$, and assume for simplicity that $\max\{z^{-1} - 1, z^{-1} - 2z\} < \overline{B} < \min\{1 + z^2, z^{-1} - z^2\}$. (For $\theta \approx 1$, this approximately implies $1 < \overline{B} < 5/4$.) A level $k+1$ agent chooses

$$C^{k+1} = C^*(C^k, B^k, z) = z\left[C^k + B^k + \left(\frac{1 - C^k}{C^k}\overline{B}\right)^+\right]$$

$$B^{k+1} = B^*(C^k, B^k, z)\begin{cases} = \overline{B} & \text{if } C^*(C^k, B^k, z) < 1, \\ \in [0, \overline{B}] & \text{if } C^*(C^k, B^k, z) = 1, \\ = 0 & \text{if } C^*(C^k, B^k, z) > 1 \end{cases}$$

If $C^k < 1$ for $k > 0$, then $B^k = \overline{B}$ and $C^{k+1} = z\left[C^k + \frac{\overline{B}}{C^k}\right] > z(1 + \overline{B} > 1$. If $C^k > 1$, then $B^k = 0$ and $C^{k+1} = zC^k$; so after T periods, where T is the smallest integer greater than $-\ln C^k / \ln z$, $C^{k+T} < 1$. That is, as $k \to \infty$, C^k alternates between values below 1 and strings of values above 1, with B^k either equal to $\overline{B}$ or 0.

Given our assumptions on z and $\overline{B}$, we can prove a stronger result: starting from initial conditions (C^0, B^0) in a neighborhood of $(1,1)$, C^k and B^k converge to a 2-cycle $(C_H, 0), (C_L, \overline{B})$ with $C_L = zC_H < 1$. First, we prove by induction that (C^k, B^k) converge to such a cycle starting from $C^0 \in (1, z^{-1})$ and $B^0 = 0$. Suppose that for some k (possibly 0), $C_k \in (1, z^{-1}), B_k = 0$. Then $C_{k+1} = zC_k < 1$ and $B_{k+1} = \overline{B}$. This in turn implies $C_{k+2} = f(C_k) := z^2 C_k + \frac{\overline{B}}{C_k}$. f is convex and decreasing on $(1, z^{-1})$ (since $f'(z^{-1}) = z^2(1 - \overline{B}) < 0$). Given our assumptions on z and $\overline{B}$, we have $f(1) = z^2 + \overline{B} < z^{-1}$ and $f(z^{-1}) = z(1 + \overline{B}) > 1$; thus, if $C_k \in (1, z^{-1})$, $C_{k+2} \in (1, z^{-1})$. Finally, our assumptions also imply $f'(1) = z^2 - \overline{B} \in (-1, 0)$, so $f'(C) \in (-1, 0)$ on the interval

$(1, z^{-1})$. This guarantees that C_k converges to the fixed point $C^H := \sqrt{\frac{\overline{B}}{1-z^2}} > 1$. Thus, level k dynamics converge to the 2-cycle $(C_H, 0), (C_L, \overline{B})$ where $C_L := zC_H < 1$.

This also implies that the system converges to the 2-cycle starting from any $C^0 \in (z, 1)$, $B^0 = \overline{B}$. Starting from such values, $C^1 = zC^0 + z\frac{\overline{B}}{C^0} \in \left(z(1+\overline{B}), z^2 + \overline{B}\right) \subset (1, z^{-1})$, $B^1 = 0$.

Now suppose $C^0 = B^0 = 1$ (as described in the main text, this implies $L0$ agents behave as in a baseline REE with $\theta = 1$, the typical assumption in macroeconomic applications of level-k reasoning). If $z \in \left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right)$, $C^1 = 2z \in (1, z^{-1})$, $B^1 = 0$. Thus, the results above imply that (C^k, B^k) converges to the 2-cycle described above. If instead $z < 1/2$, $C^1 = 2z < 1$, $B^1 = \overline{B}$, $C^2 = z[2z + \overline{B}]$. Given our assumptions, this is greater than 1. We also have

$$\begin{aligned} z[2z + \overline{B}] &< 2z^2 + z(1 + z^2) \\ &= z^{-1}[z(1+z)]^2 < z^{-1}\frac{3}{4} < z^{-1} \end{aligned}$$

So $C^2 \in (1, z^{-1})$, $B^2 = 0$, and again the result above implies (C^k, B^k) converges to the 2-cycle. Thus, for $z \neq 1/2$ but sufficiently close to $1/2$, $B^k \in \{0, \overline{B}\}$ for all k.

Finally, it is also clear that the system will converge to the 2-cycle for a specification of $L0$ behavior that is close, but not equal, to $C^0 = B^0 = 1$.

Since C^k depends on 1 even in the limiting 2-cycle, for any fixed distribution of 'cognitive depth' $\{\lambda_k\}_{k=0}^{\infty}$, we cannot have $C = \sum_{k=0}^{\infty} \lambda_k C^k = 1$ for all z, as in Nash equilibrium. One might wonder whether C converges to 1 as we shift the distribution of types to put more weight on large values of k. As $k \to \infty$, the average value of C_k converges to

$$\frac{C_L + C_H}{2} = \frac{z+1}{2}\sqrt{\frac{\overline{B}}{1-z^2}}$$

which still depends on z. When $z = 1/2$, this equals $\sqrt{\frac{3\overline{B}}{4}}$ which, given our assumption that $\overline{B} < 1 + z^2 = 5/4$, is less than 1. Thus, for generic distributions $\{\lambda_k\}_{k=0}^{\infty}$ and z close to $1/2$, $C = \sum_{k=0}^{\infty} \lambda_k C^k < 1$ will differ from 1, even if $\{\lambda_k\}_{k=0}^{\infty}$ becomes skewed towards very large values.

To be clear, the important result is not that C^k exhibits cycles, but that B^k always attains a corner solution 0 or $\overline{B}$, and never equals 1. The literature has noted that level-k dynamics can lead to (possibly explosive) oscillations in games of strategic substitutability featuring linear best response functions, and has proposed modifications, such as reflective equilibrium, which avoid this feature (García-Schmidt and Woodford, 2019; Angeletos and Sastry, 2021; Angeletos and Lian, 2023). While we study level-k dynamics in a nonlinear model, so the system exhibits a stable cycle rather than convergent or explosive oscillatory 'dynamics', the same force is at play. However, even with a 'smooth' modification of level-k implying that $C^k \to 1$, we would still have $B^k \neq 1$.