Is Utility Theory so Different from Thermodynamics?

D. Eric Smith
Duncan K. Foley

SFI WORKING PAPER: 2002-04-016

SANTA FE INSTITUTE

Is utility theory so different from thermodynamics?

Eric Smith
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

Duncan K. Foley
Department of Economics, New School University, New York, NY 10003
(April 17, 2002)

Careful examination of the axioms, and interpretation conventions, of utility theory and thermodynamics reveals that the two domains are more similar mathematically than their divergent approaches to problem solving would suggest. Their differences inhere primarily in the economic assignment of importance to initial endowments, versus the physical choice to emphasize reversible transformations. Using an analysis based on reversible transformations in economics, it is shown that utility theory can be represented as a theory of additive price potentials, and non-decrease of an additive entropy in closed systems. The standard money-metric utility is identified as a Gibbs potential for demands, and a new *contour money-metric utility* is introduced as the conjugate Helmholtz potential for prices. Examples show how the central problem-solving tools of thermodynamics apply in detail to conventional utility models, with emphasis on the concepts of reversible transformation, the equation of state, and engines.

I. INTRODUCTION

A. Different conventions; same abstractions?

Both neoclassical economics and classical thermodynamics seek to describe natural systems in terms of solutions to constrained optimization problems. The observables that link their respective theories to measurement have an obvious formal correspondence: supply and demand vectors in economics resemble the generalized energies and volumes of physical systems; prices resemble generalized temperatures and pressures. Yet there are striking differences in the ways these analogue pairs are used to solve problems in economics [18] and thermodynamics [11], and corresponding differences in the meanings attached to them.

Given the subject matters – people versus collections of atoms in a solid or gas – it might seem gratuitous to observe that the methodological approaches of thermodynamics and economics are different, and indeed the issue is not even mentioned in comprehensive introductions to economics [18,22]. The neoclassical theory of preferences [8] certainly seems constructed to make this difference look mathematically necessary, by removing from utilities almost all properties of the potentials central to thermodynamic analysis. Yet closer inspection reveals that the basic idealizations of neoclassical utility theory, and the fundamental laws of classical thermodynamics, are nearly if not strictly identical from a formal point of view. It then follows that the same structures of inference must be possible in the two fields, a consequence that this paper develops in general form and through examples.

The argument has two parts. First is a demonstration that both neoclassical economics, and classical thermodynamics, combine *theories* of states with *conventions* for attaching meanings to the state variables through transformations. The class of transformations used is different in economics from thermodynamics, in a way that is not mathematically necessary, but sharply alters the relation between theoretical prices (temperatures or pressures) and measurement. The second part of the argument is constructive. It is first shown that economic analysis makes sense when based on the thermodynamic class of transformations. From this analysis, it is then shown that all of the central quantities and theorems of thermodynamics can be constructed for general utility problems, and have natural meanings. The key to this construction is a consistent application of the method of reversible transformations in the economic context.

Given the acknowledged differences between the objects of economic and thermodynamic analysis, though, the question remains why making such a correspondence is useful. Most importantly, by separating the roles of preference axioms from those of Walrasian price interpretations, this work introduces the qualitative distinction between *irreversible* and *reversible* transformations. It shows that while economics at least since the work of Walras has tended to focus on the analysis of irreversible transformations, represented by the movement from initial endowments far from equilibrium to competitive equilibrium, nothing about either people or preferences is fundamentally Walrasian. Opening the door to analysis based on reversible transformations yields new welfare measures that generalize consumer's surplus, and provides a more complete theory of the duality of prices and allocations. Constructions from thermodynamics also provide a realization of Walras's original view of utilities as potentials for prices, and show what parts of this problem were solved by work of Fisher [12] and Negishi [20]. Furthermore, the point of view of reversible transformations opens up new and promising avenues for the analysis of economic data.

Historically, interest in mappings between economics

and physics seems to have been rooted in the hope that classical determinism would give order for free in the human world [19]. While any such hope is now long debunked (even in physics), a correspondence of utilities with physics constructs remains interesting, but for more sobering reasons. Classical thermodynamics is derived in physics from statistical properties of the most unstructured, stochastic populations of objects in nature. A certain suspension of disbelief is required to accept utility models of human behavior, on a number of grounds [1]. In the process of justifying these assumptions and approximations, it is surely valuable to know if one is assuming an underlying model that has been shown to describe supposedly far simpler systems.

At the same time, increasing numbers of stochastic models are being advanced to explain the behavior of prices and allocations. Generally these either extend or alter the assumptions of utility theory, while not being required to have limits in which it is recovered as a restricted case. Another motive for this work is therefore to show where neoclassical utility theory lies in the spectrum of thermodynamic and statistical models *as it is*, so that extensions or alterations can relate to it coherently.

B. Historical attempts at congruence

It has been argued historiographically [19], that Walras originally envisioned utilities as economic "potentials" for decision making, with the specific interpretation of prices as utility gradients, and the "physical" definition of equilibrium as the condition where supply and demand balance in the manner of forces. There were two flaws in the Walrasian correspondence, one surmountable, one fatal as originally implemented.

The surmountable flaw was that Walras sought a correspondence with rational mechanics, the physics of his day. Mechanics, however, gives no one-way rule to motions, as preference is supposed to give to economic choices, so there is no notion of the role of forces in mechanics without explicit consideration of dynamical properties like inertia. To find their equivalents in economics would have required postulating universal properties of agent response to preferences, a much more egregious assumption than anything made by utility theory alone.

This flaw was recognized and corrected by Fisher [12], who re-interpreted the Walrasian potentials in thermodynamic terms. That association will be kept in this paper, because thermodynamics has an almost-sure one-way rule for macroscopic systems: non-decrease of entropy. It is important here, though, to distinguish thermodynamic reasoning, where the one-way rule is a primitive, from more general statistical modeling, where it is a derived property of large-system limits, and loses its almost-sureness for systems of finite size. In failing to make this distinction, Fisher drew correspondences of economic agents with both thermodynamic (deterministic) and statistical (stochastic) objects, which are mutually inconsistent. This work differs from Fisher's in recognizing that a thermodynamic correspondence can be identified *without* finding the underlying statistical motivation. Indeed, it leaves as an important unanswered question whether there is any statistical theory which has neoclassical utility theory as its large-system limit.

The real significance of distinguishing between mechanical and thermodynamic concepts of utility, though, goes beyond the ability to consistently incorporate a one-way rule. In physics, it has been found that entropy *conservation* is the fundamental constraining principle for reversible transformations. Understanding its implications has led to the concepts of engine and refrigeration, and of idealized efficiencies for these processes, all of which have economic application.

The more serious flaw in the Walrasian correspondence was the assumption that cardinal forms for utilities could be constructed *a priori*, and would then have gradients that gave equal, true prices for arbitrary equilibria. Such a condition is equivalent to requiring that equilibria be minima of a global *social welfare function*, which is the sum of the utilities of all agents in an economy. This amounts to postulating a natural addition rule for cardinal utilities constructed without knowledge of the economy in which they are embedded.

In the neoclassical theory of preferences, price normalization is arbitrary [2], and utility gradients need only be parallel in equilibrium. This feature was incorporated by Negishi [20], who showed that individual competitive equilibria were minima of social welfare functions made by adding arbitrary cardinal utilities, *if* these were first multiplied by appropriate scalar weights. Multiple equilibria could not generally be obtained from the same welfare function, though, because scalar weights do not represent the full freedom to map ordinal utilities while respecting preferences.

C. Situating the neoclassical theory of value within econophysics

Much of current "statistical" economics arises from objections to the limitations of neoclassical utility theory, and it is important to emphasize that those are not the

[1] For a lucid perspective before much had been done to provide alternatives, see Ref. [1], Ch. 1. For a lookback in light of current models incorporating externalities, see Ref. [3], and references therein. For explicit development of biologically inspired models, see Ref. [10].

[2] See discussion in Ref. [8], p. 28.

concern of this paper. Roughly categorized, statistical models are of two types: Informational models [J. D. Farmer and J. P. Garahan, in preparation] seek to link informational efficiency with the entropy increase of future conditioned price sequences, rejecting the law of one price as an intrinsic characteristic of agent utilities in equilibrium. Allocational models [13,9] commit to specific trading dynamics, and assume a stochastic element in the allocation process, from which an entropy of configuration probabilities is derived.

It is essential to recognize that the potentials and entropy derived below, which encode the neoclassical constraint against decrease of utility, are already a deterministic property of agents *viewed as macroscopic systems*. It is logically impossible for entropies derived from any residual stochastic element of agent behavior to provide a deterministic constraint on each instance of that behavior. It is hoped that, once this distinction is made clear, it will help separate what is possible within conventional utility theory from what is gained (or lost) by relaxing its assumptions.

D. Mapping reversibility analysis from thermodynamics

Showing a fundamental correspondence between utility theory and thermodynamics has a necessary and a sufficient part. The necessary part is accounting for their differences in terms of conventions for interpreting prices. This step is not axiomatic, so the point will be made with a simple example in Sec. II. Once conventions have been appropriately circumscribed and their influence understood, Sec. III will show that there is a deep correspondence between the theories of states at the level of defining assumptions on preferences.

The functional correspondence of utility with thermodynamic methods will then be built term by term, using the formal results as guides. In Sec. IV, the map between expenditure functions and thermodynamic potentials will be given in general form, and its economic interpretation elaborated. In Sec. V, a detailed example will be worked out as a concrete instance of this mapping, showing how the full implications of this map carry through for a range of experiments. Sec. VI then discusses engine cycles, fundamental in all of thermodynamics, and shows that their economic interpretation is sensible and often familiar. In Sec. VII, the general process of bargaining toward the contract curve will be shown to be that of minimizing a free energy, which is the Walrasian potential for prices.

The most general utility models are somewhat less constrained than those encountered in thermodynamics, and this is why the cardinal form appropriate to global minimization can depend on the context of interactions defined by an economy. When this is the case, there is more than one possible interpretation of the economic object corresponding to physical entropy. The appropriate form is always specified when some subset of prices is stabilized, as for a small country trading a limited subset of commodities with a large world market. In the special case where utilities are linear in some commodity, there is the further possibility of defining a context-independent entropy when the economy trades that good with the rest of the world.

E. A note on language and audience

Part of the purpose of this paper is enable physicists and economists to identify their common tools and possible points of view, but to be readable it has to be presented primarily in one language or the other. The language and examples chosen will be those of economics, but to make the paper more accessible across disciplines, some terms familiar to economists but not to physicists will be defined and not just cited, and occasional justification will be provided when simplified economic models are invoked as examples.

The emphasis that will be made on reversible versus endowment-preserving transformations reflects a subtle difference of perspective between the fields, which has surprising influence on their concrete constructions. The economic emphasis on endowments, and the facilitating role of Adam Smith's invisible hand in producing equilibria, gives a particularly normative interpretation of equilibrium selection. Concretely, this is expressed in the almost exclusive use of consumer's surplus, or variations on it, as welfare measures for economies. The welfare measures most natural from the viewpoint of reversibility, in addition to being dual to consumer's surplus, have a natural "positive"[3] interpretation in terms of the net surpluses attainable from economies through voluntary exchange. It is expected that the combination of both points of view will ultimately give the most complete and versatile understanding of welfare.

II. THE ROLE OF CONVENTIONS

An elementary but sometimes under-appreciated fact is that a theory of states of equilibrium[4] is not enough

[3] In that they characterize a property intrinsic to the voluntary-exchange economy, whether it is actually extracted or not.

[4] In this paper we will use the term "equilibrium" in an economic context to refer to any Pareto-efficient allocation, that is, any configuration in which individual agents can find no mutually advantageous exchanges. As is well known, each such equilibrium allocation has an associated system of relative prices representing the common marginal rates of substitution of the agents. We will use the term "Walrasian equi-

to assign a meaning to the state variables that appear in it. There must also be some convention for measuring the values of those variables, and measurements are always made in the context of some kinds of transformations. This section will show that the choices of transformations made in economics differ from those made in thermodynamics, in a way that tries to make "prices" fundamentally different from temperatures or pressures at a practical level.

The difference is summarized in the following pair of observations: (1) There is no counterpart in physics to the way neoclassical economists attach importance to transformations respecting initial endowments. (2) Conversely, there seems to be no counterpart in economics to the importance thermodynamicists attach to reversible transformations. An emphasis on reversibility extends the definition of equilibrium to transformations in a natural way, but is generally incompatible with making interesting statements about completely closed systems. On the other hand, transformations that assign consequences to initial endowments generally cannot be made reversibly, and in admitting irreversibility, a fundamental departure is made from the uniqueness properties of equilibria.

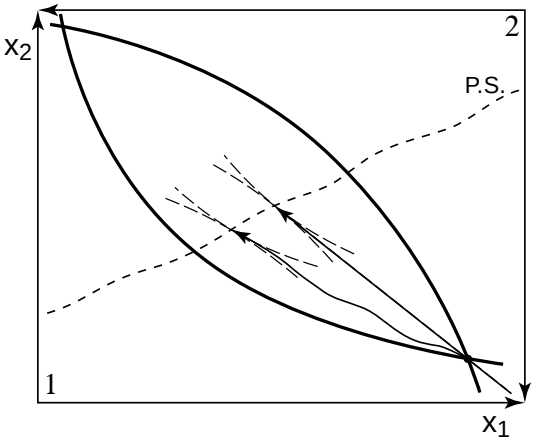

FIG. 1. Edgeworth box for a pure exchange economy with two agents. Axes are quantities of goods, x_1 and x_2. Origin for agent 1 is in the lower right, and for agent 2 in the upper left. Pareto set (short-dash) is labeled P.S. Indifference curves are the heavy black lines, and the contract curve is the subset of P.S. in the interior of the lens formed by these. Straight ray is the Walrasian map from initial endowments to the equilibrium of the First Fundamental Theorem. Wavy ray is an arbitrary, utility-improving trading history these agents could actually take. Long-dashed curves are segments of indifference curves through the equilibria attained by either path.

A. Assigning meaning to Walrasian prices

Consider the reduction of a private-ownership economy $\mathcal{E}$ [5] to no production (pure exchange), with two goods and two types of agents, each with strictly convex and insatiable preferences. As diagrammed in Fig. 1, the Edgeworth box for these agents defines a Pareto set, of points where their indifference surfaces are tangent, and a price vector (arbitrarily normalized) at each point in the Pareto set, as the normal to the tangent plane. Given whatever initial endowments define $\mathcal{E}$, the contract curve is the part of the Pareto set contained between the agents' indifference surfaces at those endowments [18].

Ref. [8] (p. 74) says of the theorem asserting sufficient conditions for the existence of at least one Walrasian equilibrium that "This fundamental theorem of the theory of value explains the prices of all commodities and the actions of all agents in a private ownership economy." The sense of "explanation", however, has a very restricted relation to measurement.

Since prices are a pure accounting tool, the only measure of the agents' wealths in $\mathcal{E}$ are the values, given a price vector, of their initial endowments that specify the economy. The existence theorem associates with that endowment at least one *wealth-preserving* equilibrium. These are, of course, the equilibria that could be reached by Walrasian auction from the initial endowments, and the existence theorem induces a correspondence from each point in the space of all initial endowments to the Pareto set.

Now, one of two alternatives is possible: either the initial endowments are in the Pareto set, or they are not. If they are, there is no contract that either agent will propose, which the other will accept in any finite amount, *at any price*. The equilibrium price is mathematically identifiable, but not measurable as a property of real exchanges, since none can occur.

Alternatively, if the initial endowments are not in the Pareto set, the Walrasian equilibrium price system in no way constrains the trades real agents can make, since either may accept any utility-improving contract proposed by the other. As recognized by Hahn and Negishi [15], there is a continuous infinity of trading histories, by which the agents may arrive at any point on the con-

librium" to denote Pareto-efficient allocations which conserve the value of some given initial endowments at the equilibrium prices.

[5] Defined in Ref. [8], p. 75.

tract curve. Since preferences may certainly be chosen so that the prices along the contract curve include an open, one-dimensional set, the equilibrium price attained after a continuous history of recontracting may take infinitely many non-Walrasian values. Thus, where prices could actually be measured as the properties of real exchange contracts, "the prices of all commodities" selected by Walrasian equilibrium need not be actual contract prices at the end of trading, or those anywhere else along its history, or even any weighted average of these.

To escape the real heart of this difficulty, it is not enough to observe that the Walrasian equilibrium price system may often be a good approximation to contract prices, for endowments in small neighborhoods of equilibria, since the axiomatic theory is concerned, above all, with what can be proven exactly. Alternatively, asserting that the function of Walrasian equilibrium prices is not to identify properties of contracts actually traded by agents, severely restricts their correspondence to the real world in a way that abstract temperatures or pressures are never restricted in thermodynamics: physical temperatures and pressures are real properties of achievable transformations, which determine the work done on or by systems.

So the key feature of temperatures, energies, pressures and volumes in physics – that they are uniquely specified, measurable properties of realizable transformations – is denied prices and allocations in Walrasian economics. If the endowment is already in the Pareto set, prices are uniquely specified but are not measurable because there are no transactions. If the endowment is not already an equilibrium, so that there will be transactions, the prices and ultimate allocation are in general impossible to specify uniquely.

B. Emphasizing reversibility rather than endowment

The preceding observations about Walrasian equilibrium prices are well-known, if frequently ignored in practical economic applications. We argue that these features of Walrasian economics are essentially methodological measurement conventions, and that they are not essential to economic theory. Walrasian auction, or any other monotonically utility-improving trading algorithm, represents an *irreversible transformation* in the language of thermodynamics. The trading history cannot be reversed because agents will generally refuse to accept trades that reduce their utility, and with suitable, standard assumptions, opposite sides of the same trade can never both be utility-improving for an agent in a given initial state.

Thermodynamics, to connect with measurement, must similarly describe transformations into or out of equilibria. However, in physics it is known that the encapsulating interface provided by temperature, energy, pressures and volumes is simply inadequate to predict the interactions of any systems not infinitesimally close to equilibrium. There is thus *no thermodynamic description* of transformations involving far-from-equilibrium system states, like the initial endowments in a Walrasian auction. (The fact that utility-improving trading histories are similarly arbitrary could be interpreted as saying the same is true in economics).

The way thermodynamics admits transformations is to require that they be *reversible*: those in which a system can be made to pass through a set of states, each arbitrarily close to some equilibrium, in both a forward and its reverse sequence. The reversible transformations in economics will be those that move agents within their indifference surfaces. Prices are uniquely and continuously defined for such movements, in a way they cannot be along the trading histories that lead from initial endowments to the contract curve.

It is obvious that systems like the foregoing two-agent economy cannot undergo nontrivial transformations that are both reversible and endowment-preserving. Thermodynamics retains a coherent treatment of equilibrium by breaking systems into components, and allowing flow into or out of the components as necessary, to keep each component arbitrarily close to an equilibrium. Neoclassical economics has taken the opposite approach, to make use of the irreversible Walrasian auction in the statements of its fundamental theorems, in order to map initial endowments to specific equilibrium allocations.

Transformations along agents' indifference surfaces are, of course, part of the analytic toolbox of standard microeconomic theory [18] The departure in this work is to assume that the *unique* interpretation of equilibrium prices is given by such transformations. Equilibrium prices are formalized, not only as the exchange rates of traded contracts, but as the encapsulating interface that makes an economic agent the equivalent of a thermodynamic system. It is this property that implies a law of one price, and there is no principle *from equilibrium* implying the same constraint on other arbitrary, utility-improving trades. The general theory of equilibrium, economic or thermodynamic, uniquely determines only this encapsulating price, and *in principle* has nothing to say about trades that cannot be entirely represented in through that interface.

C. Interpreting economies close to equilibrium

Adopting the analytical assumption that an economy is at or close to equilibrium requires some re-thinking of the interpretive substructure of economic theory. Much economics implicitly or explicitly adopts the interpretation of Hicks' *Value and Capital* [16], in which the economic time is periodized into "weeks". On Sunday night of each Hicksian week all the agents receive their endowment of commodities, and are thus on Monday morning far from equilibrium. On Monday a Walrasian market occurs, which reallocates the initial endowments (through

what we realize now is an irreversible transformation) to final commodity bundles, and the agents spend the rest of the week actually consuming those bundles. Within the framework of this parable, the actual measured transaction flows of the economy correspond to the irreversible transformations associated with the achievement of Walrasian equilibrium.

If we want to adopt the point of view of reversible transformations, it makes more sense to interpret the commodity bundles of agents as *stocks*, such as stocks of consumer durables (the food in the refrigerator, for example). The availability of well-organized markets permits agents to keep close to stock equilibrium at all times. Since agents are human beings who get hungry, wear out clothes, and generally deplete stocks, it is necessary for agents to make transactions more or less continuously to keep close to equilibrium (selling their labor-power, paying their rent, buying food, and so forth). These transactions, which generate national income, are not in this way of thinking the result of irreversible movements from far-from-equilibrium endowments to equilibrium, but the result of agents' constant effort to maintain themselves at equilibrium. Nothing like Hicks' Sunday night, in which the economy and its agents are suddenly moved to a point far from equilibrium, occurs.

The assumption that agents remain close to equilibrium at all times, and that the economy can as a result be analyzed with the concept of reversible transformations, is a strong abstraction. For example, an agent who loses her job typically feels that she has been forcibly (irreversibly) moved to a lower utility level. Real economies experience shocks (wars, revolutions, and depressions, for example) that intuitively seem to be best understood as irreversible transformations. We would like to emphasize the notion that the method of reversible transformations is best adapted to analyzing ongoing economies operating more or less normally, though the tools of reversible transformations can throw some light on the causes and magnitudes of irreversible transformations.

III. THE AXIOMS OF UTILITY THEORY AND THERMODYNAMICS

It is possible to express the abstract assumptions of utility theory in a way that both incorporates the preceding observations about transformation and measurement, and corresponds point by point with the laws defining classical thermodynamics. The defining idealizations of both fields can be seen as assumptions of *encapsulation*, *constraint*, and *preference*. In physics, these are codified in the 0^{th}, 1^{st}, and 2^{nd} laws of thermodynamics, respectively. They are not typically represented in these functional terms, but rather in terms of domain-specific phenomena (existence of temperature, energy conservation, entropy increase) through which the corresponding functions are expressed.

The defining assumptions of utility theory will be paired with their thermodynamic counterparts here, to show which phenomena implement common functions across domains. In the case of the 0^{th} law, the statement given will be slightly more general than the conventional physics form, but only to abstract away from the preferred status of time in physics, to accommodate general utility calculations in which time plays no special role.

It is nontrivial that the laws presume the existence of the two equivalent notions of *agent* and *thermodynamic* (or "macroscopic") *system*. These are given meaning through the notion of *state variables*, which will be used here to refer to either entity. The state variables are defined as the standard interface through which agents or systems interact, independent of other, arbitrarily complex details of their internal configuration. They are generally only meaningful for systems "sufficiently near" equilibrium, where such a simplified description is adequate to predict the constraints they impose on each other.

Both economic and physical state variables come in intensive/extensive variable pairs. When two systems have equal values of intensive variables, they can be aggregated to a system with the same values. The extensive variables of the aggregate system are then the sums of those from the components.

Examples of such pairs in physics are temperature/entropy, pressure/volume, chemical potential/species number density. In economics they will be the price/quantity pairs of commodity bundles. In both fields, intensive variables like prices or pressures are only uniquely associated with extensive variables like commodity stocks or volumes in the context of equilibrium. For economic agents considered in isolation, prices are defined from allocations and indifference surfaces by imagining the agents' being embedded in an economy, where those holdings are the equilibrium allocations.

A. The three laws

0^{th} **law** (*encapsulation*):

> Thermodynamic systems in equilibrium have well-defined values of some set of state variables. If two systems are brought into contact, they undergo no macroscopic changes if and only if the intensive state variables conjugate to all exchangeable quantities have the same values.
>
> Economic agents in equilibrium have well-defined prices for all commodities they can hold. If two agents are given the opportunity to trade, their holdings will remain unchanged if and only if their prices for all exchangeable commodities are equal.

The 0^{th} law is the requirement that state variables be

an adequate description of systems sufficiently near equilibrium, and that whether two systems will be in equilibrium with each other can be entirely determined from the values of their intensive state variables. Its narrower conventional statement in physics is that temperature exists for any macroscopic system in equilibrium, and that two systems placed in thermal contact exchange no heat if and only if their temperatures are equal [6].

> 1st **law** (*constraint*):
>
> Energy is conserved under arbitrary transformations of a closed system.
>
> Commodities are neither created nor destroyed simply by the process of exchange.

The importance of conservation laws is that they place constraints on the sets of possible configurations, given initial conditions. Because in physics, energy is conserved in closed systems, rather than minimized, Walras's original conception of utility as a potential energy [19] was a functional mismatch. The fact that there is no principle of energy minimization in physics, however, was only appreciated after Walras, with the discovery of the 2nd law.

> 2nd **law** (*preference*):
>
> There is a partial ordering of any configuration of state variables of a macroscopic system (given by its entropy), and transformations that decrease the entropy of a closed system almost-surely never occur.
>
> There is a partial ordering of any possible commodity bundles for an agent (given by her ordinal utility), and agents will never voluntarily accept trades that reduce the utility of their bundles.

Preference ordering is the law that identifies neoclassical agents as the equivalents of macroscopic thermal systems, and not microscopic particles. Trades that increase utility in economics, or entropy in physics, may be spontaneously taken, but if there is more than one such transformation possible, neither utility theory nor thermodynamics alone specifies a preferred choice. What happens typically depends on arbitrary details of the initial internal configuration (or chance), hence state variables alone are not enough to describe general, out-of-equilibrium transformations. The transformations in which utility or entropy are unchanged are the *reversible* transformations of either domain, and are completely described by the the changes in state variables.

A caution: the fact that both entropy and utility increase does not mean that utility *is* entropy, though in general they are related. The original formulations of entropy increase in physics were all stated explicitly in terms of *heat transfers* [11], which allow the entropy of a system that is not isolated to either increase or decrease. Understanding the economic flows that function as heat transfers, and using them to relate utility to entropy, is one main aim of the next two sections.

IV. REVERSIBLE TRANSFORMATIONS IN ECONOMIC SYSTEMS

Consider a general exchange economy with $n+1$ commodities indexed $i = 0, \ldots, n$. Commodity bundles will be denoted by vectors $\vec{x} = (x_0, \ldots, x_n)$, and x_0 taken to be the numéraire. There are m agents indexed $j = 1, \ldots, m$. Each agent has an ordinal utility function $u^j\left(\vec{x}^j\right)$ that satisfies the standard axioms, including quasi-concavity. A price system is a vector $\vec{p} = (p_0, \ldots, p_n)$. The value of a commodity bundle $\vec{x}$ at the price system $\vec{p}$ is $\vec{p} \cdot \vec{x}$.

The preferences of an agent can also be represented by the *expenditure function*:

$$e^j(\vec{p}, u^j) \equiv \min_{\vec{x}} \vec{p} \cdot \vec{x} \text{ subject to } u^j(\vec{x}) \geq u^j \tag{1}$$

The expenditure function is the mathematical dual of the utility function. The solution to the cost minimization problem that defines the expenditure function are the *Hicksian* or *compensated demand functions* $\vec{x}^{hj}\left(\vec{p}, u^j\right)$. A standard theorem in economic choice theory is that the vector of derivatives of the expenditure function with respect to prices is the vector of compensated demands:

$$\frac{\partial e^j\left(\vec{p}, u^j\right)}{\partial \vec{p}} = \vec{x}^{jh}\left(\vec{p}, u^j\right) \tag{2}$$

Reversible transformations of economies are precisely those that keep all agents at the same utility levels. (We cannot force an agent to a lower utility level through voluntary trade, and if we allow any agent to reach a higher utility level the transformation could not be reversed because we could never induce her to return to the lower utility level.) Thus the compensated demand function

[6] Explaining the special role of temperature in physics is a rather long digression. Its essence is that temperature is a characteristic scale for energies, and energy conservation follows from the symmetry under continuous time translations implicit in the definition of a temporal equilibrium. No other symmetry or conservation law is similarly priveledged by the definition of equilibrium itself. In more general systems with several such symmetries, however, at the level of classical thermodynamics, temperature becomes just another intensive state variable, and the same rules for existence and balance apply to all such variables. For a detailed treatment of the relations of time, temperature, and entropy in the context of a self-organizing dynamical system that goes somewhat beyond usual thermodynamic equilibrium, see Ref. [21].

is the natural concept though which to study reversible transformations of economic systems.

Given an arbitrary reference system of prices $\vec{q}$, the expenditure function for any agent evaluated at $\vec{q}$ and the utility level of commodity bundle x^j, $e^j\left(\vec{q}, u^j\left(x^j\right)\right)$ is a strictly increasing function of $u^j\left(x^j\right)$, and thus represents the same preferences as $u^j\left(x^j\right)$. This construction is called the *money-metric utility function at prices* $\vec{q}$, and is convenient because it expresses the utility of the agent in terms of money, that is, whatever unit of account is used to express prices, such as dollars. Until the last section of the paper we will assume that individual utilities are measured as money-metric utility functions for a given reference price system $\vec{q}$.

A. Reversible transformations in closed economies

Equation (2) shows that the sum of the expenditure functions of the individual agents at a given utility profile $u = \left(u^1, ..., u^m\right)$, $e\left(\vec{p}, u\right) \equiv \sum_j e^j\left(\vec{p}, u^j\right)$ serves as a *potential function* for reversible transformations of a closed economy:

$$\frac{\partial e\left(\vec{p}, u\right)}{\partial \vec{p}} = \sum_j x^{hj}\left(\vec{p}, u^j\right) \equiv x^h\left(\vec{p}, u\right) \tag{3}$$

The economy-wide expenditure function $e\left(\vec{p}, u\right)$ and compensated demand function $x^h\left(\vec{p}, u\right)$ encode all the behavioral information that characterizes the economy. As the price system $\vec{p}$ varies over all semi-positive values, the compensated demand function varies over all the total holdings of commodities consistent with the utility profile u. Since the expenditure functions of individual agents are convex functions of prices, so is the economy-wide expenditure function, which guarantees that the economy-wide compensated demand correspondence is one-to-one and onto. Every total holding of the economy corresponds to one and only one price system. The economy-wide compensated demand correspondence considered as a function of prices alone can be inverted to give prices as a function of total holdings: $\vec{p}\left(\vec{x}^h, u\right) = \{\vec{p} | \vec{x}^h\left(\vec{p}, u\right) = \vec{x}^h\}$.

The compensated demand function corresponds to the *equation of state* for a closed thermodynamic system. The reversible transformations of closed thermodynamic systems (the *adiabatic transformations*) are useful primarily to identify the equation of state for the system, but a closed thermodynamic system has limited capacity do any useful work. The reversible transformations of closed economic systems are interesting for basically the same reason, which is to reveal the equation of state.

The reversible transformations of a closed economy correspond to the following thought-experiment. Suppose there is a speculator, outside the economy itself, who possesses stocks of commodities, and can make small transactions with the agents of the economy involving these stocks. Imagine this speculator's gradually inducing agents to exchange one commodity for another by offering terms of trade just a little better than equilibrium prices. Such a speculator could in this way vary the total holdings of the economy without allowing agents to move to higher utility levels. In this process the equilibrium prices of the economy would change. The net vector of commodities the speculator would move into or out of the economy in such an experiment and the corresponding price changes would effectively trace out the compensated demand of the economy, and identify the behavioral surface determined by agent preferences. Of course, under the restriction of reversible transformations and a closed economy, this sequence of transactions is of no interest to a speculator, because when she reverses the transactions, she returns the economy (and her own holdings) to exactly its initial state.

This scenario, however, may not be as abstract as it might seem at first. For example, a central bank making very small changes in the interest rate on bank reserves by absorbing the excess supply or demand of reserves might be viewed as approximating such an external speculator. The scientific advantage of such natural reversible transformations is considerable, because information about the manifold of reversible transformations has powerful implications for the structure of the economy, and can be analyzed using the tools of reversible transformations developed in thermodynamics.

Thermodynamic entropy is not measurable through external manipulation of a closed system. As a corollary, there is no meaningful measurable economic quantity corresponding to thermodynamic entropy for a closed economy.

B. Reversible transformations in open economies

It is very natural in thermodynamic analysis to consider a system that is attached to a *heat reservoir* or heat bath at a fixed temperature. The system is assumed to be able to exchange effectively unbounded amounts of energy with the reservoir, so the connection to the reservoir maintains the system at the constant temperature of the reservoir. The "speculator" imagined in the last section corresponds in thermodynamic analysis to a "load", which can exchange some thermodynamic quantity with the system through reversible transformations.

The natural economic setting corresponding to a thermodynamic system connected to a heat bath is an open economy trading a good (suppose it is commodity 1) with the world market. The world market fixes the price of commodity 1, p_1, and the economy can exchange x_1 with the world market to maintain its equilibrium at that price. Now we can imagine a speculator contriving reversible transformations by making small exchanges in non-traded goods $x_2, \ldots, x_n$ with the economy that maintain all of the agents at their initial utility levels,

while the economy is connected to the world market. In the course of these exchanges two types of adjustment occur. First, the economy may exchange commodity 1 with the world market in order to maintain equilibrium consistent with the fixed p_1, and, second, the agents in the economy will be exchanging commodities to maintain equilibrium (that is, Pareto-efficiency) in the face of the changed total holdings manipulated by the speculator.

In this situation the potential function corresponding to the Gibbs potential is:

$$F = e\left(\vec{p}, u\right) - \frac{p_1}{q_1} \sum_j u^j \left(x^h \left(\vec{p}, u^j\right)\right) \tag{4}$$

The expression p_1/q_1 is a dimensionless quantity which, like a physical temperature, represents the effect of the world market in holding p_1 constant. The important property of the Gibbs potential is that its derivatives with respect to prices holding utilities constant (so that the transformations are reversible) and holding p_1/q_1 constant (respecting the connection of the economy to the world market) be the economy's total holding of commodities, x^h. It is easy to see that the quantity F fulfills this property, since the derivative of the economy-wide expenditure function gives the total holding and both p_1 and u are held constant by assumption. This Gibbs potential has a natural economic interpretation, since it is the sum of terms $e^j\left(\vec{p}, u\right) - \left(p_1/q_1\right) u^j\left(x^{hj}\right) = \vec{p} \cdot \vec{x}^{hj} - \left(p_1/q_1\right) u^j\left(x^{hj}\right)$, which are the negative of the *consumer surpluses* $\left(p_1/q_1\right) u^j\left(x^{hj}\right) - \vec{p} \cdot \vec{x}^{hj}$, the money expression of the excess of the agent's utility at the commodity bundle, adjusted for the world market valuation of traded goods over the cost of that bundle at market prices.

In thermodynamics the dual of the Gibbs potential is the *Helmholtz potential*, A, a function of the extensive variables (volumes) whose derivative with respect to those extensive variables is the negative of the intensive variables (pressures). The Helmholtz potential is related to the Gibbs potential by a *Legendre transformation*:

$$A = F - P \cdot V \tag{5}$$

where P is the (possibly vector-valued) pressure and V is the (possibly vector-valued) volume.

In the economic setting, the commodities $\bar{x} = (x_2, \ldots, x_n)$ correspond to the volumes, and the Helmholtz potential is:

$$\begin{aligned} A\left(p_0, p_1, \bar{x}\right) &\equiv e\left(\vec{p}, u\right) - \frac{p_1}{q_1} \sum_j u^j \left(x^h \left(\vec{p}, u^j\right)\right) - \bar{p} \cdot \bar{x}^h \\ &= p_0 x_0 + p_1 x_1 - \frac{p_1}{q_1} \sum_j u^j \left(x^h \left(\vec{p}, u^j\right)\right) \end{aligned} \tag{6}$$

where $\bar{p} = (p_2, \ldots, p_n)$.

To see that this quantity satisfies the properties of the Helmholtz potential, consider its variation for a given commodity $j > 2$ holding the other commodities except x_0 and x_1 constant. We have $\delta A = p_0 \delta x_0 + p_1 \delta x_1$ because p_1/q_1 and u are being held constant in a reversible transformation, in an economy connected to the world market. But from the budget constraint $\delta\left(p_0 x_0 + p_1 x_1 + \bar{P} \cdot \bar{x}\right) = 0$, so that $p_0 \delta x_0 + p_1 \delta x_1 = -\bar{p} \cdot \delta \bar{x}$ so that:

$$\frac{\partial A}{\partial \bar{x}} = -\bar{p} \tag{7}$$

In economic terms, we isolate x_0 in this derivation because it represents the budget constraint (since we are taking x_0 as the numéraire) and x_1 because it is the commodity traded with the world market, and hence its quantity is endogenous.

In this setting, there is a meaningful economic quantity corresponding to thermodynamic entropy, which is $\sum_j u^j\left(x^{hj}\right) - q_1 x_1$. The economic quantity corresponding to thermodynamic temperature is p_1/q_1, which is being held constant by trade with the world market. Changes in the utility component of this quantity correspond to irreversible transformations increasing utility (and this quantity), and changes in the $q_1 x_1$ component reflect flows of the traded good (corresponding to heat in the thermodynamic system) with the world market reservoir. This entropy-like quantity can change (either rise or fall) in the course of a reversible transformation, but only through trade. The conservation of the traded commodity in exchange shows that this entropy change must be offset by a corresponding entropy change in the rest of the world, just as an entropy gain or loss of a thermodynamic system must be balanced by a corresponding loss or gain of the reservoir.

C. Economic-thermodynamic correspondences

In this section we have demonstrated one of the main claims of this paper, that in the context of reversible transformations, there is a functional correspondence between the formalisms of economics and thermodynamics. Each of the key thermodynamic concepts can be interpreted consistently (and intuitively) in the economic context. These correspondences reveal that thermodynamic entropy and economic utility are related, though not identical. Economic utility corresponds precisely to that component of thermodynamic entropy whose change arises from irreversible transformations.

V. A WORKED EXAMPLE

There are two specific purposes for the following example. One is to show, in some detail, why the previous interpretation of reservoir-traded goods as energy and entropy variables is both functionally general and non-arbitrary. It adds intuition to what economic "energy"

means and does, and shows why its function is distinct from that of the x_1 contribution to the entropy.

The other purpose is to develop a further interpretation of the Gibbs and Helmholtz potentials that holds in physics: these are the thermodynamic generalizations of mechanical potentials, as measures of the height of a system above equilibrium. Importing this functionality into economics will give the interpretation of the Helmholtz potential as a welfare measure, dual to the measure defined by consumer's surplus. Because it implies a relation between a system at and away from equilibrium, it will require that the definitions of these potentials be extended from individual indifference surfaces to the whole space of demands. Doing so will lead to a new money-metric utility, dual to the utility defined from the expenditure function. For any agent, it is parametrized by one free demand curve, which may then be chosen to ensure that arbitrary collections of individual utilities aggregate sensibly to describe community welfare.

A. Debt-linear utility

The model considered here describes valuation of one-period returns from investment in a risky, dividend-bearing asset, versus the cost of risk-free borrowing, and an individual production function on held capital. The only important respect in which it is not general is that the ordinal utility will be linear in one reservoir-traded good, which gives a unique, preferred definition to entropy and temperature. Sec. VII shows how the construction generalizes when quasi-linearity is not assumed.

There are three commodities: $(x_0, x_1, x_2) \equiv (M, -D, N)$. M is capital (arbitrarily in the form of money), N is the number of shares of a dividend-paying asset, and D is debt service owed at the end of a period[7] of length δt.

For reversible transformations, an agent's (differential) budget constraint defines prices p for shares, and interest rates r for borrowing, as

$$\delta M = -p_N \delta N + \frac{1}{r\delta t}\delta D. \tag{8}$$

Thus $(p_0, p_1, p_2) \equiv (1, 1/r\delta t, p_N)$ in the notation of Sec. IV.

Once her allocation is set, the wealth change of an agent over the period is

$$\Delta W \equiv Nd - D + \phi(M), \tag{9}$$

where d is the payment of that realization of the dividend process. $\phi(M)$ is the production function of held capital. (For instance, if M is used to make car and house payments needed to hold a certain job, $\phi(M)$ is the period income in excess of those expenses.) ϕ will be assumed monotone increasing and concave.

If all allocation decisions are based only on the wealth change over a period (9), the utility model is a pure dividend-discount model. Expectations over future price fluctuations, or other factors used in valuing real assets, are omitted here for simplicity.[8]

A simple ordinal utility, quadratic in N, is produced [2] by the CARA[9] cardinal utility of wealth change

$$u_{\text{card}} \equiv -\exp\{-\Delta W/\nu\}, \tag{10}$$

and normal distribution of dividend payments with mean $\langle d\rangle \equiv \bar{d}$ and variance $\left\langle \left(d-\bar{d}\right)^2 \right\rangle \equiv \bar{d}^2\sigma^2$. ν is the agent's risk tolerance, measured in units of M. The ordinal utility is the expected cardinal utility over the dividend distribution,

$$u = -\exp\left\{-\left[N\bar{d}\left(1-\frac{N\bar{d}}{2\nu}\sigma^2\right) - D + \phi(M)\right]/\nu\right\}, \tag{11}$$

and up to monotone transformation is equivalent to the debt-linear *certainty equivalent of wealth*, denoted with the special symbol

$$\mathcal{U} \equiv N\bar{d}\left(1-\frac{N\bar{d}}{2\nu}\sigma^2\right) - D + \phi(M). \tag{12}$$

The expenditure function corresponding to the utility (12) is

$$e^j(\vec{p},\mathcal{U}) = M + p_N N - \left[\phi(M) + N\bar{d}\left(1-\frac{N\bar{d}}{2\nu}\sigma^2\right) - \mathcal{U}\right]/r\delta t, \tag{13}$$

where the dependence of M and N on $\vec{p}$ will be derived below.

[7] Debt service is defined as the interest owed per period on all borrowed capital, and does not include principle repayment. Since borrowing is not assumed to take place at any one interest rate, principle is only repaid by "selling" it back to the lender (or to anyone else) in a voluntary exchange.

[8] This simplification is valid for sufficiently rare information shocks relative to the discounting horizon. In that case, the importance of infrequent price changes can always be made negligible relative to accumulated per-period wealth in the interim. An alternative application is to measure the arbitrage opportunities afforded by wrong expectations, by asking what capital can be extracted from agents who always expect zero price change, and then find themselves in a world where prices have changed.

[9] Abbreviation for Constant Absolute Risk Aversion.

B. From local to global one-way rules

Agents accept only trades for which

$$\delta\mathcal{U} = \delta N \bar{d}\left(1 - \frac{N\bar{d}}{\nu}\sigma^2\right) + \delta M \frac{\partial\phi}{\partial M} - \delta D \geq 0. \quad (14)$$

For a single agent, this one-way rule is equivalent to non-decrease of entropy. This is not enough, however, to imply that utility is *equal to* entropy. In thermodynamics, the entropy of subsystems also has a natural addition rule under aggregation. Aggregated subsystems may be able to take configurations that they could not take in isolation, through newly possible *heat flows*, but the local one-way rules must, through the entropy addition rule, impose a global constraint on configurations of the aggregate.

The natural addition rule for utilities in this system is uncovered by separating the flow from the stock variables in the combination

$$S \equiv \mathcal{U} + D = N\bar{d}\left(1 - \frac{N\bar{d}}{2\nu}\sigma^2\right) + \phi(M). \quad (15)$$

The stocks M and N are the natural generalized energy and volume coordinates for an agent, independent of the time period δt, which it is often desirable to regard as arbitrary. Now, not just because $\mathcal{U}$ has units of D, but because D is actually *traded*, additivity of D extends naturally to S. We note also that Eq. (14) is equivalent to

$$\delta S \geq \delta D. \quad (16)$$

To define the proper behavior of entropy, and its relation to heat flows between small economies and reservoirs, it is easiest to start by considering all agents first as a closed system, and then separating the reservoir explicitly as a large (and featureless) subsystem. This extended economy is the collection of agents indexed $j \in 1, \ldots, m$, with consumption bundles $\left(M^j, -D^j, N^j\right)$, and utilities $\mathcal{U}^j$ incorporating production functions ϕ^j and risk tolerances ν^j. Since it is closed all risk-free lending is done internally, and for any configuration,

$$\sum_{j=1}^{m} D^j = 0. \quad (17)$$

Extending this addition rule to S^j makes it an extensive state variable under aggregation, with sum

$$S_{\mathrm{TOT}} \equiv \sum_{j=1}^{m} S^j = \sum_{j=1}^{m}\left(\mathcal{U}^j + D^j\right) = \sum_{j=1}^{m}\mathcal{U}^j. \quad (18)$$

Since each $\delta\mathcal{U}^j \geq 0$ in any trade, it follows that $\delta S_{\mathrm{TOT}} \geq 0$ for all voluntary transformations of the system, even though any given S^j may increase or decrease.

S^j and S_{TOT} are candidates for agent j's, and total system entropy, respectively. The step that identifies them as the correct candidates is explicit identification of the reservoir that defines temperature.

C. A world market for debt

Under reversible transformation, the budget constraint (8) becomes

$$\delta M = -p_N \delta N + \frac{1}{r\delta t}\delta S. \quad (19)$$

Equivalently, combining Eq. (8) and Eq. (14) at $\delta\mathcal{U} = 0$ gives

$$\left(r\delta t - \frac{\partial\phi}{\partial M}\right)\delta M = \left[\bar{d}\left(1 - \frac{N\bar{d}}{\nu}\sigma^2\right) - (r\delta t)\, p\right]\delta N, \quad (20)$$

so the interest rate sets

$$r\delta t = \frac{\partial\phi^j}{\partial M^j} \quad (21)$$

for any agent j.

The natural reservoir to consider in this problem fixes r. We retain one agent, say $j = 1$, as the "small economy", and make all other agents so risk averse that they hold no shares of the asset ($\nu^j \to 0, j \in 2, \ldots, m$). When m is taken large, these constitute a world market for debt. The speculator (or load) is conveniently instantiated as the issuer of the untraded good (shares).

The thermodynamic 1st law for a system connected to a reservoir and load is written [11]

$$\begin{aligned}\delta U &= -\delta W + \delta Q \\ &= -p\delta V + T\delta S. \end{aligned} \quad (22)$$

U (not $\mathcal{U}$!) would be the system's internal energy, δW the increment of work done by her on the load, and δQ the increment of heat absorbed from the reservoir. The first line of Eq. (22) is general, but not generally useful unless the transformation is reversible, in which case δW and δQ have the expressions in the second line, in terms of pressure (p), volume (V), temperature (T), and entropy (S). For transformations out of equilibrium, pressure need not be defined as a state variable, and in general $\delta Q < T\delta S$; that is, entropy increases more than is accounted for by the inward heat flow.

Mapping Eq. (19) formally onto Eq. (22), $U \leftrightarrow M$, $V \leftrightarrow N$, $p_N \leftrightarrow p$ (share price is pressure), $T \leftrightarrow 1/r\delta t$, and S from Eq. (15) is indeed the entropy. *Capital* is the variable corresponding to internal energy, and its conservation constrains the set of possible configurations, as desired. The heat flow into the small economy is then $\delta D/r\delta t$ (the part of capital bought with debt), and for irreversible transformations where $\delta\mathcal{U} > 0$, Eq. (16) verifies that $\delta Q < T\delta S$. The following subsections will show that this mapping indeed respects all other functional relations, as well.

For general reversible transformations, Eq. (20) then relates the instantaneous share price and interest rate to holdings as

$$\frac{p_2}{p_1} = (r\delta t)\, p_N = \bar{d}\left(1 - \frac{N\bar{d}}{\nu}\sigma^2\right) = \mathcal{F}(x_2)\,. \tag{23}$$

Eq. (23) is called the single agent's *equation of state*, a relation between the non-numéraire price components, and a function $\mathcal{F}$ of the (extensive) compensated demand x_2.

D. Construction of the free energies

The original goal of thermodynamics was to generalize the mechanical notion of a *potential energy*, to some consistent representation of the maximum work that can be extracted from a system, which can undergo irreversible as well as reversible transformations. Clearly this is also what economists have sought, to preserve the deterministic features of equilibrium, while freeing their methodology from the the excessive dynamical assumptions required for predictions in mechanics.

This generalization is accomplished by the notion of *free energy*. When a system undergoes an irreversible transformation, the potential work extractable from it decreases. To assign quantity to that decrease by means of the reversible transformations (which one can constrain by the equation of state (23)), the system can alternatively be brought reversibly to the same (or an equivalent) final condition, *only* if a reservoir is first connected to compensate for the extracted work. Thus all free energies are defined explicitly in terms of the reservoir assumed.

Nothing in the following constructions demands that the reservoir hold a single price fixed; in particular the construction of the contour money-metric utility below could be done under much more general models of the world market. However, in both economics and physics, price (temperature) regulation is both an easily described and a widely applicable function of reservoir exchange, so only that case will be discussed here explicitly. Because it is often interesting to consider reservoirs that hold more than a single price fixed, it is necessary to have thermodynamic potentials that describe work extraction in the contexts of either fixed-volume or fixed-pressure boundary conditions. The Gibbs potentials and Helmholtz potentials are thus both required to give a complete dual description.

The expenditure function (13) in this example may be written

$$e(\vec{p}, \mathcal{U}) = F + \frac{1}{r\delta t}\mathcal{U}, \tag{24}$$

as in Eq. (4). Consistent with the identification $T = 1/r\delta t = p_1/q_1$, one recognizes that $\mathcal{U}$ is already in money-metric form, with $q_1 = 1$.

From the definition (15) of S, the form of F is

$$F = M + p_N N - \frac{1}{r\delta t}S. \tag{25}$$

An important thermodynamic relation of the entropy to the pressure is recovered from Eq. (15) and Eq. (23) for the quantity S,

$$\left.\frac{\partial S}{\partial N}\right|_M = (r\delta t)\, p_N = \frac{p}{T}, \tag{26}$$

and because M is fixed by Eq. (21) in terms of $r\delta t$,

$$\left.\frac{\partial F}{\partial p_N}\right|_{r\delta t} = N. \tag{27}$$

Relation (27) could also have been obtained by taking the most general variation of F, and then using the definition of equilibrium prices as normal to indifference surfaces to cancel terms from δM using Eq. (19). Using this method, the variation of the Legendre transform $A \equiv F - p_N N$ is computed as:

$$\begin{aligned}\delta A|_{r\delta T} &= \delta\left(M - \frac{1}{r\delta t}S\right)\Big|_{r\delta T} \\ &= -p_N \delta N \\ &\equiv -\delta W.\end{aligned} \tag{28}$$

As required,

$$p_N = -\left.\frac{\partial A}{\partial N}\right|_{r\delta t}. \tag{29}$$

From the last line of Eq. (28), it is clear that A measures the potential capital extractable by voluntary exchange from the agent/reservoir system by the speculator, and can be interpreted as an intrinsic measure of the distance of arbitrary economic initial conditions from equilibrium.

The free-energy/intrinsic-welfare interpretation maps a standard interesting problem in physics to a standard interesting problem in economics. The welfare of a collection of agents who, for whatever reason, cannot trade their way to equilibrium, is the potential profit of a market maker who introduces the necessary trade service.

In this context, suppose the small economy consists of two agents, both borrowing from a world market at rate r, and the speculator is replaced with a Market Maker (MM) who trades exclusively with the agents. MM's defining characteristic is that she is infinitely risk averse, and so maintains zero inventory of shares in any round of trading.

The resulting economy is diagrammed in Fig. 2, and is identical to the physical problem of a piston between two pockets of gas, not initially at the same pressures, but sharing a fixed volume. In physics one asks what is the maximal work that can be extracted by an experimenter coupled to the piston, who cannot herself absorb volume from the system.

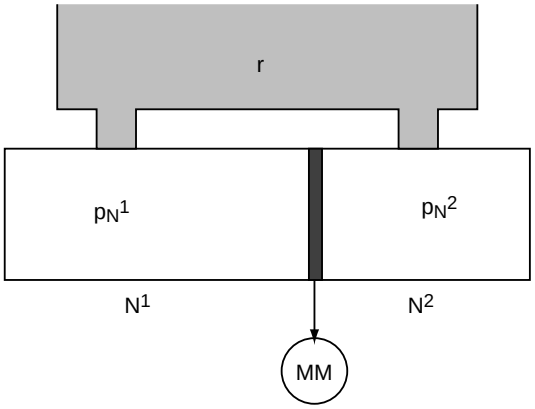

FIG. 2. Profit extraction by a Market Maker (MM), from two agents with different share prices ${p_N}^1$ and ${p_N}^2$, borrowing at fixed rate r. Share holdings N^1 and N^2 correspond to the volumes of chambers on either side of the MM "piston".

The agents hold share numbers N^1 and N^2, with $N^1 + N^2 \equiv N$ fixed. They have equilibrium prices ${p_N}^1$ and ${p_N}^2$, in general different. MM offers least-favorable prices to each agent in a sequence of incremental trades, moving them finite distances along their indifference surfaces, while extracting capital

$$\Delta M_{MM} = \int \left({p_N}^1 - {p_N}^2\right) dN^1 = \int {p_N}^1 dN^1 + {p_N}^2 dN^2$$
$$= -\int d\left(A^1 + A^2\right), \tag{30}$$

the change in their combined Helmholtz potentials. The lower limit of integration defines the initial condition for N^1, and the upper limit is whatever N^1 gives the agents equal price (pressure), after which there is no more potential for capital extraction (work).

The interesting economic point is that the potentials in Eq. (30) are integrals of information obtainable entirely from the equation of state, which may in principle be mapped independently of this process. To see how this plays out, we first note that the specific form of the Helmholtz free energy of agent $j \in 1, 2$ is

$$A^j = M^j - \frac{1}{r\delta t}\left[N^j \bar{d}\left(1 - \frac{N^j \bar{d}}{2\nu^j}\sigma^2\right) + \phi^j\left(M^j\right)\right]. \tag{31}$$

It is convenient to parametrize share holdings relative to the values that will ultimately produce equilibrium prices at the current ν^j:

$$N^1 \equiv \frac{\nu^1}{\nu^1 + \nu^2} N - N',$$
$$N^2 \equiv \frac{\nu^2}{\nu^1 + \nu^2} N + N'. \tag{32}$$

Supposing that non-equilibrium prices arose because agent 1 previously equilibrated her holdings at a risk aversion ${\nu^1}_0 < \nu^1$, and then that risk aversion changed, the equation of state is as shown in Fig. 3. (ν^2 was always as it is now.)

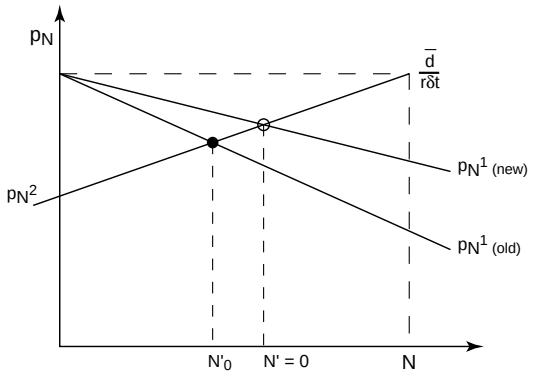

FIG. 3. Prices versus N' for the agents of Fig. 2. Agent 1 was originally more risk averse, leading to price function ${p_N}^1{}_{(\mathrm{old})}$, and an allocation equilibrium at ${N'}_0$. Agent 1's current, reduced risk aversion produces price curve ${p_N}^1{}_{(\mathrm{new})}$, and equilibrium allocation $N' = 0$, against the price function ${p_N}^2$ for agent 2.

The combined free energy for the two agents,

$$A^1 + A^2 = \left(M^1 + M^2\right) - \frac{1}{r\delta t}\left[\phi^1\left(M^1\right) + \phi^2\left(M^2\right)\right]$$
$$- \frac{N\bar{d}}{r\delta t}\left(1 - \frac{N\bar{d}}{2\left(\nu^1 + \nu^2\right)}\sigma^2\right)$$
$$+ \frac{\bar{d}^2\sigma^2}{2r\delta t}\left(\frac{1}{\nu^1} + \frac{1}{\nu^2}\right){N'}^2, \tag{33}$$

is minimized at the equal-price point $N' = 0$, and its initial value is given by $N' = {N'}_0 > 0$. (Because both M^j are fixed by r, the first line of Eq. (33) is a constant that does not enter into optimization decisions.) The maximum reduction in Eq. (33) is

$$\Delta\left(A^1 + A^2\right) = -\frac{\bar{d}^2\sigma^2}{2r\delta t}\left(\frac{1}{\nu^1} + \frac{1}{\nu^2}\right){N'}_0^2, \tag{34}$$

whereas direct integration of the first line of Eq. (30), with prices given by the equation of state (23), gives

$$\Delta M_{MM} = \frac{\bar{d}^2\sigma^2}{2r\delta t}\left(\frac{1}{\nu^1} + \frac{1}{\nu^2}\right){N'}_0^2, \tag{35}$$

as desired.

VI. MULTIPLE RESERVOIRS AND ENGINE CYCLES

Unless the small agents' disequilibrium is constantly replenished, the market maker's activities in the last section are more relevant to one-shot arbitrage than to the creation of a sustained pattern of activities. A more interesting question is what can be extracted by a small speculator operating between heterogeneous *reservoirs*, which cannot trade with each other directly. This becomes a problem for the speculator when the good which she can readily exchange (say, shares) is one for which

the reservoirs do not have disparate prices, or in which they do not trade at all.

Exactly this problem in physics led to the concept of *engine cycles*. The typical situation is that free energy is in principle available from reservoirs with different temperatures. Yet as we now know, temperature is the energy-price of the entropy which flows as heat. Because the entropy of physical thermodynamics arises from uncontrolled degrees of freedom in statistical mechanics, the speculator (macroscopic experimenter) always finds herself in the position of being limited to interactions which cannot directly trade entropy [10].

The importance of engines, in physics or economics, is that the speculator can still extract the excess value of this system, and that she can extract an unlimited amount of it (for infinite reservoirs), even using a small market to mediate the exchange between heat flows and work (world-market-traded, and untraded, goods). The essential reason is that the small market can be guided through *repeatable cycles*, as long as only reversible transformations are used. These cycles can be repeated indefinitely many times, each cycle extracting fixed wealth, as long as the temperature differences are maintained.

An engine is instantiated in this example by allowing a single agent (or small country) to interact alternately with capital reservoirs (world markets) at different interest rates r_1 and r_2.[11] The speculator once again issues shares N, and functions as a mechanical load.

The simplest engine cycle to analyze is the *Carnot cycle*. It comprises four reversible transformations, shown as a price-volume diagram for the agent's state in Fig. 4. First, the speculator lowers prices p_N under conditions of constant rate r_1, and the agent borrows from the capital market to acquire $N_2 - N_1$ shares. For whatever reason, the agent is then rendered unable to borrow, but the speculator continues to lower p_N, inducing demand for additional $N_3 - N_2$ shares bought by depleting the agent's private capital M. Because ϕ is concave, this increases the lending rate at which she would be in equilibrium to $r_2 > r_1$. Supposing borrowing is then restored, but now from a capital market at r_2, the speculator can raise p_N to buy back $N_3 - N_4$ shares at reduced cost (the dividend is less attractive to the agent relative to risk-free lending). To close the cycle, borrowing is then suspended again, and the speculator buys back shares $N_4 - N_1$, by raising p_N to its initial level. N_4 is chosen relative to N_{1-3} so as to restore the agent's private capital, and hence equilibrium rate, to their original values as well.

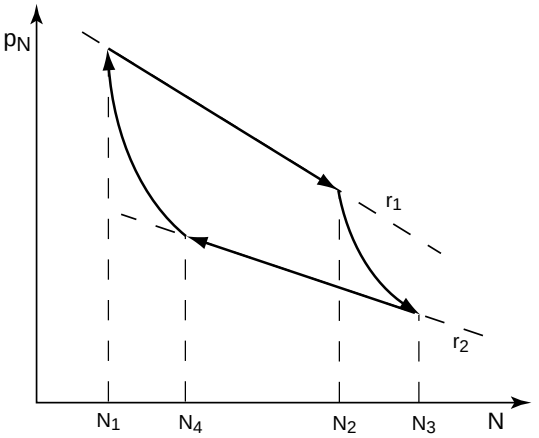

FIG. 4. The Carnot cycle state diagram for a single market agent. Cycle starts at N_1, r_1, and legs are followed (and numbered) in the direction of the arrows.

The device of suspending borrowing while r changes is invoked to limit discussion to two lending reservoirs (r_1 and r_2 only). However, the profit extracted by the speculator is just a consumer's surplus, the area inside the closed curve of Fig. 4. The area inside *any* such closed curve could be arbitrarily well covered by sufficiently small tiles of Carnot type, so the mathematical conveniences taken here imply no loss of generality.

An important property of engines (monetary or energetic) is immediately apparent from Fig. 4. Though there may be unlimited capital held by the world markets, only a fraction of it can be extracted by the speculator through voluntary exchange, and the amount extracted in each cycle is strictly less than the capital the speculator can induce the agent to borrow on leg one. (This may be important if the single agent or small country is credit-constrained.) An elementary principle of accounting, called *Carnot's theorem*, shows that the most total capital is extracted by reversible cycles, and that that fraction depends only on the properties of the reservoirs.

The essence of the proof is that the agent must acquire debt to borrow capital on leg one, and in a closed cycle, all of that debt must be sold back on leg three. This can be written as a cyclic integral of the debt change $\oint dD = 0$, where there happens to be no debt flow on legs two and four. Since the price of debt is determined

[10] Seen carefully, this is actually a definition of macroscopic measurement, and the situation in statistical physics is *exactly* that of the speculator in the untraded good. It is the category of measurements of a system with many degrees of freedom that *defines* which of them are "uncontrolled". Precisely the energy in these degrees of freedom remains expressed as temperature inequality, even if all other sources of work mediated by macroscopic transformations (such as pressure differences) have been equalized. The essential point is not *how many* degrees of freedom are in a statistical system, but only that there are *more* degrees of freedom than those that can be controlled by the "judges" [14] defined by measurements.

[11] One can imagine, for instance, that the world market discounting horizon is much shorter than the speculator's and can drift slowly. Then succeeding time intervals with different interest rates define perfectly sensible temporary equilibria to the world and the small country, but appear as distinct reservoirs to the speculator.

by the world market on leg one, the most capital that can be collected by the speculator is exactly this *heat flow* from the reservoir, denoted

$$Q_1 \equiv \int_1 \frac{1}{r_1 \delta t} dD. \tag{36}$$

Similarly, no less capital will sell back the debt at r_2 than the heat flow on leg three

$$Q_3 \equiv \int_3 \frac{1}{r_2 \delta t} dD. \tag{37}$$

For any cycle that returns the single agent to her original holdings, $\oint dM = 0$, and so the upper limit on the speculator's profit is $Q_1 - Q_3 = (1 - r_1/r_2)\, Q_1$.

Carnot's theorem is the statement that entropy flux must be conserved in closed cycles, and it arises precisely because entropy (here, debt) is a good which the speculator cannot trade. Therefore, all entropy which a finite agent takes on in the course of an engine cycle must be traded back to the reservoirs for the cycle to be repeatable. This can in fact be used to derive the unique form of entropy given the equation of state [12]. The derivation is given in App. A, and recovers the form of S used above. The steps in the proof of Carnot's theorem are filled in there as well.

VII. VOLUNTARY-EXCHANGE WELFARE MEASURES AND IRREVERSIBILITY

The only changes in entropy from reversible transformation come from the compensated demand component x_1, so the form of utility used in Sec. IV, and up to now in the worked example, has been arbitrary. The money-metric utility $\mathcal{U}$ of the example, however, has the property that A defines the potential for capital extraction from the agent, under *both reversible and irreversible* transformations. This section shows how that property can be preserved for arbitrary (well-behaved) preferences and reservoirs, by monotone transformation of the ordinal utility to a suitable cardinal form. We use the example to understand what properties the cardinal form must have.

The instantiation of the Helmholtz potential (6), corresponding to Eq. (25), for the two agents $j = 1, 2$ of Fig. 2, is

$$A^j = M^j + \frac{1}{r\delta t}\left(-D^j\right) - \frac{\mathcal{U}^j}{r\delta t}. \tag{38}$$

Under reversible transformation, both $\delta\mathcal{U}^j = 0$, but the budget constraint for each agent sets $\delta\left(M^j - D^j/r\delta t\right) = -{p_N}^j \delta N^j$. If the market maker of Fig. 2 had not coupled to the system, and the agents had been allowed to trade to equilibrium themselves, we would have had both $\delta M^j = 0$, both $\delta D^j = 0$, but $\delta\mathcal{U}^j > 0$, for at least one of $j = 1, 2$. However, the utility $\mathcal{U}^j$ itself satisfies

$$\frac{\partial\left(\mathcal{U}^j/r\delta t\right)}{\partial\left(M^j, -D^j, N^j\right)} = \left(1, \frac{1}{r\delta t}, {p_N}^j\right), \tag{39}$$

so for *either type* of transformation $\delta A^j = -{p_N}^j \delta N^j$.

Because the market maker never accumulated shares, the agents irreversibly exchange the same sequence of δN^j as they exchanged reversibly. Because, by the equation of state (23), ${p_N}^j$ depends only on N^j (and not $\mathcal{U}^j$), they also follow the same sequence of ${p_N}^j \delta N^j$, and stop trading at the same holdings. Thus, through *any* combination of reversible and irreversible trades leading to a Pareto optimum, the total $\Delta\left(A^1 + A^2\right)$ is the same.

The reason the agents measure their own welfare gain as the capital that could have been extracted from them otherwise, is that $\mathcal{U}$ is itself a Helmholtz potential for prices. Just as the standard money-metric utility [22], constructed from the expenditure function, defines a Gibbs potential from arbitrary (well-behaved) preferences, it is possible to construct a dual, *contour money-metric* utility which is a Helmholtz potential, for any (similarly well-behaved) preferences. $\mathcal{U}$ of the example happens to be of this form.

The contour money-metric utility $\mu\left[\vec{c}; \vec{x}\right)$ for arbitrary preferences is constructed in App. B. Whereas the standard money-metric utility is parametrized by a reference price $\vec{q}$, the contour money-metric utility is parametrized by a continuous demand curve $\vec{c}(\lambda)$, and it is easy to understand why this must be. Where a Gibbs potential depends on prices, its dual should depend on demands. But whereas every price vector is associated with a point on any indifference surface through the compensated demand function, there is no such mapping for demand vectors. It is necessary to introduce a contour that functions as a *connection* on the space of goods, mapping some demand on any indifference surface to a particular demand on any other indifference surface, through translation of λ.

It is proved in App. B that, when the contour for a pair of agents $j = 1, 2$ coupled to an arbitrary reservoir is chosen to be their Pareto set (suitably generalized to account for the reservoir), and $\mu\left[\vec{c}; \vec{x}\right)$ taken as the definition of $(p_1/q_1)\, u^j$ in Eq. (6), the Helmholtz potentials A^j satisfy the following theorem:

Theorem 1 *For ordinal u^1 and u^2 representing strictly convex, insatiable preferences, $A^1\left(\vec{x}^1\right) + A^2\left(\vec{x}^2\right)$ has the following properties:*

1. *$A^1 + A^2 = const.\ \forall\left(\vec{x}^1, \vec{x}^2\right)$ in the generalized Pareto set.*
2. *$\delta\left(A^1 + A^2\right) \leq 0$ for all voluntary trades involving only $\vec{x}^1$, $\vec{x}^2$, and the reservoir, with strict inequality*

[12] For the corresponding derivation using the ideal gas in physics, see Ref. [11], p. 46-62.

for initial conditions not in the generalized Pareto set (δ denotes change in value).

3. *The maximum reduction in $A^1 + A^2$ from voluntary trades, leaving the reservoir-untraded goods $\bar{x}^1 + \bar{x}^2 = \bar{x}_{\mathrm{TOT}}$, is the maximum profit an external trader can extract from the agent-reservoir system through voluntary trading.*

4. *This maximum extractable profit is strictly decreasing under any voluntary trades involving only $\vec{x}_1$ and $\vec{x}_2$ and the reservoir, and vanishes on the generalized contract curve.*

The theorem is proved for two agents, but because the compensated demand function associates a unique bundle with each agent for any price vector, the result immediately extends to A^1 and A^2 representing arbitrary aggregations, as long as the agents in each aggregate are understood to be in equilibrium with each other.

Thus, there is a utility construction for an arbitrary economy, which behaves sensibly as a potential in the space of demands. Its sum is a uniquely defined social welfare function, of which *all* of the Pareto optima are local minima. Since they form a continuous set, they must also be degenerate, so that the gradient of the welfare function is orthogonal to the Pareto set in sufficiently small neighborhoods of it, and vanishes only on the equilibria. Using $\mu\left[\vec{c};\vec{x}\right)$, each agent measures her own welfare, relative to any point in the Pareto set, as the amount of money she fails to lose if she can bargain to that point, rather than to some other Pareto optimum to which she is indifferent.

The special feature of quasilinear utilities is that, for reservoirs trading the linear good, μ depends on the contour by at most a constant offset, independent of allocations or trading histories. Then it does not matter who else is in the economy, and the forms of utility, Gibbs, and Helmholtz potentials depends only on initial allocations. This corresponds in all details with the usual case in physics, which may be seen as reflecting the special nature of time in dynamical systems. More generally in economics, the prices to which agents may bargain themselves through utility-increasing trades need not be those to which an external speculator could bring them, though the Helmholtz welfare remains uniquely defined for each agent, given the economy in which she is embedded.

A. Time continuity and short-term debt

The ability to define welfare functions for agents, in terms only of their preferences and trading reservoir, is an important simplification. It is so universal in physics that it is often implicitly associated with the assumptions of thermodynamics *a priori*. This has probably contributed to the perception that thermodynamics is not a suitable paradigm for economic description, because its assumptions are too constrained. It is therefore worth asking what, about the nature of time, makes this simplification so common, and whether such universality may apply to some economic problems.

The essential point is that each commodity in an economy is usually assumed to define an independent basis vector, with only the discrete topology in relation to any other commodity. An example where this may ignore relevant structure is that of short-term debt markets.[13].

Specifically, define a short-term debt market as one in which the ability to make loans is so unrestricted as to have no intrinsic, shortest time scale. This is presumably a good approximation to the function of many very liquid markets in which the properties of debt securities are not highly volatile. For such markets, the only scaling of debt service that has a so-called *continuum limit* (one defining no intrinsic short-time scale for contracts) is $D \propto \delta t$. Similarly, the only nontrivial, nonsingular dependence of an intertemporal utility that can describe arbitrarily fine-grained lending decisions is a sum of terms linear in the D for each time interval. By monotone transformation, any such utility can be reduced to quasilinear form in D.

B. Relation to Negishi's social welfare functions

The ability to construct a social welfare as a sum of agent utilities, whose gradient vanishes at any specific Pareto optimum, has been noticed before [20]. The limitation of that work is that the cardinal form of the utilities is assumed specified beforehand, in which case a different set of weights must multiply each agent utility in the welfare function, to recover each equilibrium. Building welfare functions from the contour money-metric utility is, in a sense, the obvious generalization of that construction, given two observations.

First, only preferences are primitive, so there is no reason to limit transformations of an initially specified u to scalar multiplication. Second, for strictly convex preferences, every point in the Pareto set intersects a unique indifference surface for each agent, so the gradient may be scaled independently at each optimum. The multiplier in the first line of Eq. (B6) is the necessary scale factor. It is not constant in general, defining rather a nonlinear monotone transformation of u.

VIII. DISCUSSION

Parts of the correspondence that we have put forward, between utility theory and thermodynamics, have been

[13] There may be other cases where a continuous topology on goods is appropriate. cf. Ref. [8] regarding rental of a truck continuously parametrized by its mileage, etc.

recognized before. What we have called the encapsulating function of state variables, and the equivalence of the budget constraint to energy conservation, have been elaborated in the context of macroeconomics [4]. The similarity of the one-way rules enforced by utility and entropy increase has also been noticed, and the isomorphism demonstrated between the utility- and entropy-representation problems [6].

The idea of measuring the distance of an allocation above optimality, in terms of excess goods at equivalent satisfactions, is represented in the Coefficient of Resource Allocation [7]. It differs from the Helmholtz potential derived here, in being defined through a single-price exchange of finite quantities of goods, as is natural from a Walrasian perspective, rather than through reversible transformation. Within their respective methodologies, though, both are monetary measures, compatible with voluntary trade, of the efficiency with which resources are used to satisfy the desires of a community.

What does not seem to have been appreciated is that the correspondence of utility theory to thermodynamics defines a whole consistent methodology, and not just a set of analogies. In some cases, this has merely limited the scope of the conclusions: Where the representation problem was shown to be isomorphic for entropy and utility [6], the quantities themselves were only regarded as "similar", and the fact that there is a constructive, *non-identity* relation between them was not observed.

In others, it has led to mixing of robust associations with model-dependent interpretations that do not generalize, or even to the invention of unmeasurable quantities to fill gaps in proposed mappings. Ref. [17] correctly identifies the necessary correspondence of a neoclassical agent to a macroscopic system (and recognizes Fisher's error in not having done this consistently [12]), and calls utility an "analogon" of entropy, stopping short of any explicit functional mapping. It then goes on, though, to interpret the value $\vec{p}\cdot\vec{x}$ (our notation) as the equation of state, in order to map specific variables to those in the equation of state of the physical ideal gas. A similar attempt to map to this particular form (which is not even universal in physics) has led to the invention of a "productive content", with no econometric definition, to fill the place of temperature in that equation [5].

We have developed the risky-dividend model in some detail, to show how the same associations are made under the constraints of a single, consistent framework of interpretation. While there is in general an equation of state for an agent, specified by the compensated demand function, its form comes from the specification of the economic model, and need not describe any particular system in physics. More generally, we have tried to give substance to the distinction between functional equivalence and analogy, by applying all of the fundamental physical constructions to a pre-existing, conventional economic form.

Acknowledgments

The authors would like to thank Sam Bowles, Philip Mirowski, Makoto Nirei, and especially Doyne Farmer for many helpful discussions, ideas, suggestions, and references. DKF acknowledges the McKinsey Corp. for support while at SFI.

APPENDIX A: DERIVATION OF THE ENTROPY FROM THE ENGINE CYCLE

The differential budget constraint, integrated over any closed cycle, implies that

$$\oint dM = -\oint p_N\, dN + \oint \frac{1}{r\delta t} dD = 0. \tag{A1}$$

From the equation of state (23), the work integral along the first leg satisfies

$$\int (r\delta t)\, p_N dN = \int dN\, \bar{d}\left(1 - \frac{N\bar{d}}{\nu}\sigma^2\right), \tag{A2}$$

a function only of the endpoint shares N_2 and N_1. Thus the capital paid to the speculator on this leg is

$$\begin{aligned}\Delta M_1 &\equiv \int_{N_1}^{N_2} p_N dN \\ &= \frac{1}{r_1\delta t}\left[S_N(N_2) - S_N(N_1)\right],\end{aligned} \tag{A3}$$

where

$$S_N(N) \equiv N\bar{d}\left(1 - \frac{N\bar{d}}{2\nu}\sigma^2\right). \tag{A4}$$

Using Eq. (A2), plus Eq. (21) along the second leg, relates S_N to ϕ as

$$\begin{aligned}\int_{N_2}^{N_3} \frac{\partial\phi}{\partial M} p_N dN &= S_N(N_3) - S_N(N_2) \\ &= -\int_{N_2}^{N_3} \frac{\partial\phi}{\partial M} dM \\ &= \phi(r_1) - \phi(r_2).\end{aligned} \tag{A5}$$

The capital paid back by the speculator on the third leg satisfies

$$\Delta M_3 = -\frac{1}{r_2\delta t}\left[S_N(N_3) - S_N(N_4)\right], \tag{A6}$$

and on the fourth leg, the constraint between S_N and ϕ is

$$S_N(N_1) - S_N(N_4) = \phi(r_2) - \phi(r_1). \tag{A7}$$

It immediately follows that $S(N,r) \equiv S_N(N) + \phi(r)$ is a state variable held constant on legs two and four, and

whose change on leg three must exactly reverse that on leg one:

$$S_N(N_3) - S_N(N_4) = S_N(N_2) - S_N(N_1). \tag{A8}$$

The form of S from Eq. (15) is thus recovered, with $\phi(r) \equiv [\phi(M) \mid \partial\phi/\partial M = r\delta t]$. Zero borrowing on legs two and four is zero heat flow, with $\delta Q = T\delta S$.

The relation (A8), with Eq. (A1), then gives the efficiency relation

$$\oint p\, dN = \left(1 - \frac{r_1}{r_2}\right) Q_1 \tag{A9}$$

on the speculator's profit over a cycle, where Q_1 is defined in Eq. (36)

Eq. (A9) is equivalent to the statement $\oint dD = 0$. Any greater profit at the same Q_1 requires $\oint dD > 0$, with debt acquired at r_1 exceeding that cancelled at r_2. Carnot's theorem is the statement that this cannot be achieved through voluntary exchange. Proof: for any convex preference relation on $(M, -D, N)$, there is an inverse function

$$D = \mathcal{D}(M, N, \mathcal{U}). \tag{A10}$$

Supposing that debt service owed is an undesirable good,

$$\left.\frac{\partial \mathcal{D}}{\partial \mathcal{U}}\right|_{N,M} < 0 \tag{A11}$$

everywhere. If the cycle integral is expressed in terms of initial and final values as $\oint dD \equiv D_f - D_i$, greater than Carnot efficiency requires $D_f - D_i > 0$ at the same M and N, implying by Eq. (A11) that $\mathcal{U}_f - \mathcal{U}_i < 0$ and contradicting preference. QED.

APPENDIX B: THE CONTOUR MONEY-METRIC UTILITY

Let $\vec{c}(\lambda)$ be any parametrized curve in the space of demands with the property that, for any ordinal utility u representing some given agent's preferences,

$$\left.\frac{\partial u}{\partial \vec{x}}\right|_{\vec{x}=\vec{c}(\lambda)} \cdot \frac{d\vec{c}}{d\lambda} \neq 0, \tag{B1}$$

so that $\vec{c}$ intersects every indifference surface exactly once, transversally. A value of the parameter λ is assigned to every point in the commodity space by u. Denoted $\lambda_u(\vec{x})$, it is the point on $\vec{c}$ with the same utility as $\vec{x}$:

$$u(\vec{c}(\lambda))|_{\lambda=\lambda_u(\vec{x})} \equiv u(\vec{x}). \tag{B2}$$

Where it simplifies notation, $\vec{c}(\vec{x}) \equiv \vec{c}(\lambda_u(\vec{x}))$ will be used below. $\vec{c}$ will be called a *connection* for the utility u on the commodity space.

Given a point $\vec{x}$, connection $\vec{c}$, and some arbitrary reference value λ_0 (chosen independently for each agent) a *contour money-metric utility* is defined by

$$\mu[\vec{c}; \vec{x}] \equiv \int_{\lambda_0}^{\lambda_{\mathcal{U}}(\vec{x})} \vec{p}(\vec{c}) \cdot d\vec{c}. \tag{B3}$$

(The square bracket indicates dependence as a functional; regular parenthesis dependence as a function.) The line element is defined as $d\vec{c} \equiv (d\vec{c}/d\lambda)\, d\lambda$, and $\vec{p}(\vec{c})$ is the equilibrium price at $\vec{c}$:

$$\vec{p}(\vec{c}) \equiv \left.\frac{\partial u/\partial \vec{x}}{\partial u/\partial x^0}\right|_{\vec{c}}. \tag{B4}$$

By the transversality condition (B1), there is a unique λ at every indifference surface, related to the line element and price by

$$\vec{p}(\vec{c}) \cdot \frac{d\vec{c}}{d\lambda} = \frac{du/d\lambda}{\partial u/\partial x^0|_{\vec{c}}}. \tag{B5}$$

It follows immediately that

$$\begin{aligned} \frac{\partial \mu[\vec{c}; \vec{x})}{\partial \vec{x}} &= \frac{\partial u/\partial \vec{x}|_{\vec{x}}}{\partial u/\partial x^0|_{\vec{c}(\vec{x})}} \\ &= \left(\frac{\partial u/\partial x^0\big|_{\vec{x}}}{\partial u/\partial x^0|_{\vec{c}(\vec{x})}}\right) \vec{p}(\vec{x}), \end{aligned} \tag{B6}$$

where $\vec{p}(\vec{x})$ is the equilibrium price vector for u at $\vec{x}$. Assuming that capital is a desirable good everywhere, $\mu[\vec{c}; \vec{x})$ simply provides a rescaling of the price vector at any $\vec{x}$ by a $\vec{c}$-dependent ratio.

1. Edgeworth boxes with reservoirs

A reservoir for two goods defines a constraint

$$\delta\left(x^0 + p_r x^1\right) = 0 \tag{B7}$$

on all trades in which it participates, with p_r independent of volume traded. It will be convenient to introduce a balance-of-payments coordinate $a \equiv x^0 + p_r x^1$, which is not changed by reservoir exchange, and some linearly independent coordinate which measures explicitly that exchange, such as $b = x^0 - p_r x^1$, and to write

$$\vec{x} \equiv (a, b, \bar{x}), \tag{B8}$$

where as in Sec. IV, $\bar{x} \equiv (x_2, \ldots, x_n)$. The value of $\vec{x}$ at $\vec{p}$ then decomposes as

$$\vec{p} \cdot \vec{x} \equiv a + \bar{p} \cdot \bar{x}. \tag{B9}$$

Writing the connection components in the same basis,

$$\vec{c} \equiv (c_a, c_b, \bar{c}), \tag{B10}$$

the contour money-metric utility takes the form

$$\mu\left[\vec{c};\vec{x}\right] \equiv \int_{\lambda_0}^{\lambda_{\mathcal{U}}(\vec{x})} dc_a + \bar{p}\left(\vec{c}\right)\cdot d\bar{c}. \tag{B11}$$

To prove Theorem 1, it is convenient to generalize the construction of an Edgeworth box to the case of two agents interacting with a reservoir, by projecting the ordinal utility of $\vec{x}$ onto a utility for the components $(a,\bar{x})$ conserved in interactions with the reservoir. The projection works because, in coordinates $(a,b,\bar{x})$, u is still concave, but no longer monotonic in b. For any equilibrium supported at price p_r, there is a $b_u\left(a,\vec{x}\right)$ for which

$$\left.\frac{\partial u}{\partial b}\right|_{a,b_u(a,\bar{x}),\bar{x}} \equiv 0. \tag{B12}$$

A *reservoir effective utility*

$$u_r\left(a,\bar{x}\right) \equiv u\left(a,b_u\left(a,\bar{x}\right),\bar{x}\right) \tag{B13}$$

then satisfies

$$\begin{aligned}\left.\frac{\partial u_r}{\partial a}\right|_{a,\bar{x}} &= \left.\frac{\partial u}{\partial a}\right|_{a,b_u(a,\bar{x}),\bar{x}},\\ \left.\frac{\partial u_r}{\partial \bar{x}}\right|_{a,\bar{x}} &= \left.\frac{\partial u}{\partial \bar{x}}\right|_{a,b_{\mathcal{U}}(a,\bar{x}),\bar{x}}.\end{aligned} \tag{B14}$$

The arguments of u_r will simply be denoted $(a,\bar{x}) \equiv \tilde{x}$ below.

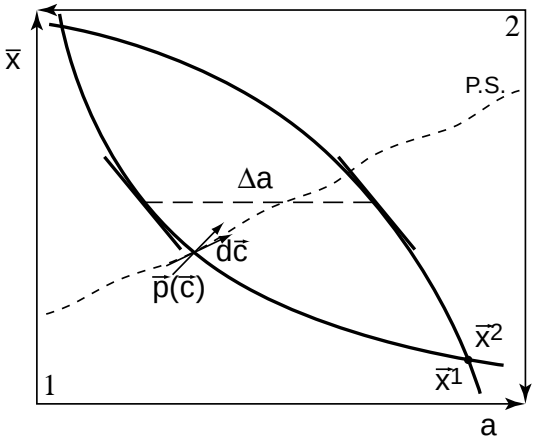

FIG. 5. The Edgeworth box for projected allocations $(a,\bar{x})$, in the presence of reservoir lending. Indifference curves are of ${u_r}^1$, ${u_r}^2$, and the associated Pareto set (short-dashed) is labeled P.S. Maximal profit potential from initial allocations $\vec{x}^1$, $\vec{x}^2$ (shown in projection) is the long-dash segment labeled Δa.

The Edgeworth box for two agent bundles $\tilde{x}_1$, $\tilde{x}_2$ in equilibrium with the reservoir, and effective utilities ${u_r}^1$, ${u_r}^2$, is shown in Fig. 5. Under arbitrary interaction of only those two agents and the reservoir, there is a fixed quantity of $a^1 + a^2 \equiv a_{\rm TOT}$, and of $\bar{x}_1 + \bar{x}_2 \equiv \bar{x}_{\rm TOT}$. In particular, the Pareto set for the three-component economy (the agents and reservoir) projects onto a single curve in $\tilde{x}$.

The contour money-metric utility (B3) is a coordinate-independent *length* of the contour $\vec{c}$, defined in terms of the agent's preferences and a normalization convention for prices ($p^0 = 1$). The sum of these lengths for the agents becomes an invariant function of the Pareto optima if each chooses as her contour some segment of the Pareto set ending on her current indifference surface. This will be denoted by referring to the whole Pareto set in Fig. 5 as $(c_a,\bar{c})$, and letting the agent's utility and starting point (denoted here λ_0^j) indicate that she measures her holdings along it. (Only the differentials dc_a, $d\bar{c}$ appear in the definition of μ, and they are opposite for the two agents.) The remaining component for each agent j is her $b_{u^j}\left(c_a,\bar{c}\right)$. Since the sum $b_{u^1} + b_{u^2}$ is not fixed in general, the pair of curves $\left(c_a, b_{u^1}\left(c_a,\bar{c}\right),\bar{c}\right)$, $\left(c_a, b_{u^2}\left(c_a,\bar{c}\right),\bar{c}\right)$ will be called a *generalized* Pareto set for this economy.

With this choice of connections, and the contour money-metric utility defining the expression $(p_1/q_1)\,u^j$ in Eq. (6), the Helmholtz potential takes the form

$$A^j\left(\vec{x}^j\right) \equiv a^j - \mu\left[\vec{c};\vec{x}^j\right], \tag{B15}$$

and satisfies Theorem 1 of Sec VII.

Proof: By construction of the Edgeworth box, for any $(\tilde{x}_1,\tilde{x}_2)$ in the Pareto set of ${u_r}^1, {u_r}^2$, $A^1 + A^2$ depends only on the endpoints of the two utility integration contours:

$$\begin{aligned}A^1 + A^2 &= a_{\rm TOT} - \int_{\lambda_0^1}^{\lambda_0^2} dc_a + \bar{p}\left(\vec{c}\right)\cdot d\bar{c}\\ &\equiv A_{{\rm TOT},CC}.\end{aligned} \tag{B16}$$

This proves point *1.*

For general $(\tilde{x}_1,\tilde{x}_2) = \tilde{x}_{\rm TOT}$, $A^1 + A^2$ differs only by the length of the segment of the Pareto set between the indifference surfaces ${u_r}^1\left(\tilde{x}^1\right)$ and ${u_r}^2\left(\tilde{x}^2\right)$ (a positive quantity by construction); hence

$$A^1 + A^2 = A_{{\rm TOT},CC} + \int_{\lambda_{u^1}\left(\vec{x}^1\right)}^{\lambda_{u^2}\left(\vec{x}^2\right)} dc_a + \bar{p}\left(\vec{c}\right)\cdot d\bar{c}. \tag{B17}$$

From Equations (B6) and (B17) it follows that, for any trades involving only the agents and reservoir,

$$\delta\left(A^1 + A^2\right) = -\frac{\delta u^1}{\partial u^1/\partial x_0^1|_{\vec{c}(\vec{x}^1)}} - \frac{\delta u^2}{\partial u^2/\partial x_0^2|_{\vec{c}(\vec{x}^2)}}. \tag{B18}$$

For voluntary trades $\delta u^1 \geq 0$, $\delta u^2 \geq 0$, with strict inequality of one utility for any transformations not in the generalized Pareto set. For convex preferences x_0 can be chosen to represent a desirable good everywhere, so both $\partial u^j/\partial x_0^j$ are positive, proving *2.*

For any trades involving an external agent, if the residual holdings $\bar{x}^1 + \bar{x}^2 \neq \bar{x}_{\rm TOT}$, there is no mechanism

to convert the bundle extracted to cash, using only the agents and reservoir. Therefore the maximum profit extractable is the

$$\Delta a = \max\left[a_{\rm TOT} - \left(a^1 + a^2\right)\right] \tag{B19}$$

that occurs on the initial indifference curves, where

$$\left.\frac{\partial \bar{x}^1}{\partial a^1}\right|_{{u_r}^1} = \left.\frac{\partial \bar{x}^2}{\partial a^2}\right|_{{u_r}^2} \tag{B20}$$

and $\bar{x}^1 + \bar{x}^2 = \bar{x}_{\rm TOT}$. From Eq. (B14) it follows that the gradients of u^1 and u^2 are parallel, so that Δa is $\max\left[a_{\rm TOT} - \left(a^1 + a^2\right)\right]$ in the full configuration space, constrained by $\bar{x}_{\rm TOT}$ and equilibrium with the exchange reservoir. Since motion along indifference curves preserves $\mu\left[\vec{c}, \vec{x}\right)$, the differential trades between the initial and final configurations satisfy

$$\delta\left(A^1 + A^2\right) = \delta\left(a^1 + a^2\right) \tag{B21}$$

giving *3*.

Finally, the expression for the change in a^j in Eq. (B19), from any change in initial configuration, is

$$\delta a^j = \frac{\delta {u_r}^j}{\partial {u_r}^j / \partial a^j} - \bar{p} \cdot \delta \bar{x}^j, \tag{B22}$$

where $\bar{p}$ is the same for both agents. Since voluntary trades increase either ${u_r}^1$ or ${u_r}^2$, and $\delta\bar{x}^1 + \delta\bar{x}^2 = 0$, it follows that $\Delta a \leq 0$ in Eq. (B19) for voluntary changes in initial conditions, with strict inequality when the initial state is not in the generalized Pareto set, giving *4*. QED.

The quantities in the proof are diagrammed in Fig. 5. It is noteworthy that, for a general pair of utilities, the contract curve is not coincident with the curve along which the difference Δa is measured, and that curve need not occur at the same values of $\bar{x}^1$ for different pairs of indifference surfaces.

2. The contour entropy

From Eq. (B15) and the form (6), it is clear that $\mu\left[\vec{c}; \vec{x}\right)$, which depends explicitly on the reservoir price p_r, defines $\left(p_1/q_1\right) u^j$. Letting q_1 be an arbitrary reference price for x^1, independent of the particular state of the reservoir, this gives the expression for the entropy conjugate to p_r:

$$S\left(\vec{x}\right) = \frac{q_1}{p_r}\left(\int_{\lambda_0}^{\lambda_{\mathcal{U}}(\vec{x})} \vec{p}\left(\vec{c}\right) \cdot d\vec{c} - p_r x^1\right). \tag{B23}$$

When preferences are quasilinear, and x^1 is chosen as the linear good,

$$\frac{\partial \mathcal{U}/\partial x^1}{\partial \mathcal{U}/\partial x^0} = p_r \tag{B24}$$

fixes $\partial\mathcal{U}/\partial x^0$, and by Eq. (B5), renders

$$S\left(\vec{x}\right) = q_1\left(\frac{u\left(\vec{x}\right) - u\left(\lambda_0\right)}{\partial u/\partial x^1} - x^1\right) \tag{B25}$$

independent of the contour $\vec{c}$, up to a constant offset. In that case, $u\left(\vec{x}\right) / \left[\partial u/\partial x^1\right] - x^1$ is also independent of x^1.

[1] Arrow, K. J., and F. Hahn, "General Competitive Analysis", Holden-Day, San Francisco, CA, 1971.

[2] Bouchaud, J. P., and M. Potters, "Theory of Financial Risks", Cambridge U. Press, New York, 2000, p. 109

[3] Bowles, S., and H. Gintis, Walrasian economics in retrospect, *Q. J. Econ.* **115** (2000), 1411-1439

[4] Bródy, A., K. Martinas, and K. Sajo, An essay in macroeconomics, *Acta Oeconometrica* **35** (1985), 337-343

[5] Bryant, J., A thermodynamic approach to economics, 36-50, Butterworth and Co., 1982.

[6] Candeal, Juan C., Juan. R. De Miguel, Esteban Induráin, and Ghanshyam B. Mehta, Utility and Entropy, *Economic Theory* **17** (2001), 233-238.

[7] Debreu, G., The Coefficient of Resource Allocation, *Econometrica* **19** (1951), 273-292.

[8] Debreu, G., "Theory of Value", Yale Univ. Press, New Haven, CT, 1987

[9] Dragulescu, A., and V. M. Yakovenko, Statistical mechanics of money, *Eur. Phys. J.* B**17** (2000), 723-729.

[10] Farmer, J. D., Market force, ecology, and evolution, http://www.arxiv.org/adapt-org/9812005.

[11] Fermi, E., "Thermodynamics", Dover, New York, 1956.

[12] Fisher, I., Mathematical Investigations in the Theory of Value and Prices (doctoral thesis), *Transactions of the Connecticut Academy* Vol. IX, July 1892

[13] Foley, D. K., A statistical equilibrium theory of markets, *J. Econ. Theory* **62** (1994), 321-345

[14] Gell-Mann, M., and S. Lloyd, Information Measures, Effective Complexity, and Total Information *Complexity* **2** (1996), 44-52.

[15] Hahn, F., and T. Negishi, A Theorem on Nontatonnement Stability, *Econometrica* **30** (1962), 463-469.

[16] Hicks, J. R., "Value and capital; an inquiry into some fundamental principles of economic theory", Clarendon Press, Oxford, England, 1946.

[17] Lisman, J. H. C., Econometrics, Statistics and Thermodynamics, The Netherlands Postal and Telecommunications Services, The Hague, Holland, MCMIL, Ch. IV.

[18] Mas-Collel, A., M. D. Whinston, and J. R. Green , "Microeconomic Theory", Oxford Univ. Press, New York, 1995.

[19] Mirowski, P., "More Heat than Light", Cambridge Univ. Press, Cambridge, England, 1989.

[20] Negishi, T., Welfare Economics and the Existence of an Equilibrium for a Competitive Economy. *Metroeconomica* **12** (1960), 92-97.

[21] Smith, E., Carnot's theorem as Noether's theorem for thermoacoustic engines, *Phys. Rev.* E**58** (1998), 2818-32.

[22] Varian, H. R., "Microeconomic Analysis", third edition,
Norton, New York, 1992, ch. 7 and ch. 10.