arXiv:2607.17015v1 [physics.soc-ph] 19 Jul 2026

Alignment of a Total Automation Economy

David McAllester

July 21, 2026

Consumption is the end purpose of production.

Adam Smith

History is driven by economics.

Historical Materialism — Karl Marx

Government of the people, by the people, and for the people, must not perish.

Abraham Lincoln, The Gettysburg Address

Abstract

We consider economic theory from the perspective of a total automation economy, one with no human involvement in production either in manufacturing or in management. One can naturally ask whether a total automation economy is fundamentally a centrally planned economy or, alternatively, whether efficiency demands decentralization into local decisions by competing agents — agentic production. A soviet economist, Leonid Kantorovich, developed linear programming as a method companies or governments can use to optimize production. Ironically, he is also generally credited with showing that the most efficient production is achieved through decentralization — a free market economy with competing agents. I will call Kantorovich's result the agentic theorem and review it in detail here. This analysis appears to reveal a alignment vulnerability under which agentic management of the economy diverges from human values as agents deveop motives that are "internal" to the economy and disconnected from prices derived from human consumption markets.

1 Introduction

Here we consider Kantorovich's dualization theorem [Kan39] from the perspective of a total automation economy — one where no human is used in the production of goods and services. The statement of the theorem assumes that the goal of industrial production is maximizing consumer value defined as maximizing consumption weighted by market price. One can think of market pricing as a form of democracy where people "vote" on what is to be produced simply by purchasing what they want. The ability of the economy to respond to the purchases of consumers can be viewed as a form of alignment of the economy with human value. The fundamental question asked here is how to maintain faithful alignment to human value (defined by purchases) in a total automation economy.

There are valid fairness issues to be address in treating purchases as reflective of social welfare. Obviously rich people have more "voting power" (purchasing power) than poor people. The needs of poor people are more pressing and should perhaps be given even more weight than the needs of rich people (instead of the reverse). Also, people in different states of health have different significance of their needs. A final issue is that different people simply have different fundamental values perhaps stemming from differences in their fundamental natures. Market pricing would seem to better reflect human value if we assume that all consumers have the same universal basic income. However, here we put the fairness issues aside and simply assume a certain mathematical model in which the objective is "simply" to maximize consumption weighted by price.

We focus here on an alignment vulnerability in this model involving the lack of pricing for internal commodities — commodities produced and consumed within manufacturing processes and never offered to consumers. Commodities not offered to consumers cannot be assigned market prices. There appears to a possibility of runaway internal process under the control of AI agents pursuing profits under shadow prices.

2 Formal Set Up

The formal set up here is due to Kantorovich [Kan39], who won a Nobel Prize in economics (1975) for formulating and analyzing this model. We assume a set of N commodities indexed by i. We assume K processes indexed by k and let z_k be a rate at which process k runs. Each process consumes commodities as inputs and produces commodities as outputs. For commodity i and process k we write $\mathrm{In}_{k,i}$ for the rate commodity i is consumed by process k when process k runs at a unit rate. Similarly $\mathrm{Out}_{k,i}$ denotes the rate of production of i when k is run at unit rate. We assume an industrial "endowment" which is a supply of natural resources (for example land) and let Endow_i be a rate at which commodity i is produced as an endowment. Given a set of processes running at given rates z_k we have a net production of rate x_i for commodity i given by

$$x_i(z) = \mathrm{Endow}_i + \sum_k z_k(\mathrm{Out}_{k,i} - \mathrm{In}_{k,i})$$

We say that a set of rates $z_k \geq 0$ is feasible if for each commodity i we have $x_i(z) \geq 0$. We cannot consume more of commodity i in production than is made available by the endowment plus production. A commodity i with $x_i(z) = 0$ will be called an internal commodity — internal commodities are used in production but are not provided to consumers. Engines might be used in making cars but never offered to consumers. A commodity will be called external if it is not internal (and hence offered to consumers in a market).

We define the purpose of the the economy to be that of maximizing consumption weighted by price. To define prices we invoke the Arrow-Debreu theorem [AD54] which implies that for a given net production $x_1, \ldots, x_n$ there exists a market equalibrium price vector $p_1, \ldots, p_n$ such that when human consumers are offered the commodities in quantity $x_1, \ldots, x_n$ at prices $p_1, \ldots, p_n$ the market clears — consumers buy exactly the amount produced.[1] A fundamental issue is that p_i is not defined for internal commodities which are not offered to consumers. This is a wrinkle in the optimization problem that is not present in the general analysis of KKT conditions for constrained optimization. However, as explained below, we can adopt the convention that $p_i = 0$ for any internal commodity i.

We note below that the failure to establish human value for internal commodities is a potential alignment vulnerability. Note that if the economy were ever to run with no external commodities then the alignment objective becomes zero everywhere and the optimization can go in any direction consistent with never producing external commodities. However, alignment through maximization of $\sum_i p_i x_i$ has nice properties. The economy should respond to human value as measured by market pricing.

[1] The market prices are sensitive to the income distribution among consumers. Since the economy is running without human involvement it is natural to assume that each consumer has the same universal basic income. But nothing in the analysis done here relies on that assumption.

3 A Central Planning Optimality Condition

We will work with differential changes in production rates $dz_1, \ldots dz_k$. Differentiating the definition of $x_i(z)$ we have

$$dx_i = \sum_k (\mathrm{Out}_{k,i} - \mathrm{In}_{k,i}) dz_k$$

A process k will be called inactive if $z_k = 0$ and called active otherwise. For feasible production rates z we say that a differential update direction $dz_1, \ldots, dz_K$ is feasible if for each inactice process z_i we have $dz_i \geq 0$ and for each internal commodity i we have $dx_i \geq 0$. A feasible update direction must not violate the feasibility restrictions on z. However a feasible update direction might convert an internal commodity to an external commodity or vice-versa or convert an inactive process to an active process or vice-versa.

> **Optimality**: We say that a feasible setting for production rates $z_1, \ldots, z_k$ is (first order) optimal if for any feasible update direction $dz_1, \ldots, dz_K$ we have $\sum_i p_i dx_i \leq 0$.

If market prices are always positive (not typically required — people can get paid to take garbage) this optimality condition is sound in that if the condition fails then improvement in the objective is possible. However, it is incomplete in that even if the condition holds improvement might still be possible through the conversion of an internal commodity to an external commodity at a non-zero market price. In spite of the incompleteness of this condition, and the unusual requirement that market prices are always positive, one can ask whether this optimality condition can be localized into profit motives for individual agents of production.

4 The Agentic Theorem

Conversion of the central objective into agentic profit motives is done using the KKT (Karush-Kuhn-Tucker) conditions on constrained optimization [KT51, Kar39]. We take the objective function to be $\sum_i p_i x_i$ with p_i defined on all commodities i but set to zero on internal commodities. Applying the KKT conditions we introduce a Lagrange multiplier $\lambda_i \geq 0$ for each constraint $x_i \geq 0$ and a Lagrange multiplier $\mu_k \geq 0$ for each constraint $z_k \geq 0$. We set $\lambda_i = 0$ on external commodities and $\mu_k = 0$ on active processes.

We have that only one of p_i and λ_i can be non-zero. This is not just complementary slackness and is not a property of constrained optimization generally. In the general case of constrained optimization the "price vector" p is just the gradient of the (fixed) objective function which can be anything. Complementary slackness is just the property that λ_i can be nonzero only if the associated constraint is active. Here we have **defined** $p_i = 0$ for internal commodities. The KKT conditions will apply to the price vector as if that vector is a well-defined gradient of the objective — ignoring the fact that an arbitrarily small change in z can move a price discontinuously from zero to being market-determined.

Before applying the KKT conditions we first note that

$$\nabla_z x_i(z) = \sum_k (\mathrm{Out}_{k,i} - \mathrm{In}_{k,i}) \delta_k$$

Here we have that $\nabla_z x_i(z)$ is a K-dimensional vector and $(\mathrm{Out}_{k,i} - \mathrm{In}_{k,i})$ is the partial derivative of $x_i(z)$ with respect to z_k — a scalar. This scalar becomes a K-dimensional vector when multiplied by δ_k — the vector whose kth component is 1 and all other components are zero.

For maximizing $\sum_i p_i x_i$ with constraints of the form $x_i(z) \geq 0$ and $z_k \geq 0$ the KKT condition is

$$\begin{aligned} 0 &= \sum_i p_i \nabla_z x_i(z) + \sum_i \lambda_i \nabla_z x_i(z) + \sum_k \mu_k \delta_k \\ &= \sum_k \delta_k \left(\mu_k + \sum_i (p_i + \lambda_i)(\mathrm{Out}_{k,i} - \mathrm{In}_{k,i}) \right) \end{aligned}$$

The sum over k is summing over orthogonal vectors so this equation yields that for each k we have

$$\mu_k + \sum_i (p_i + \lambda_i)(\mathrm{Out}_{k_i} - \mathrm{In}_{k,i}) = 0$$

For every active process we have that $\mu_k = 0$ and hence all active processes must operating at zero profit under the pricing $(p_i + \lambda_i)$. For each inactive process we can satisfy the equation by setting μ_k provided that $(p_i + \lambda_i)(\mathrm{Out}_{k,i} - \mathrm{In}_{k,i}) \leq 0$. We now have the following localization theorem.

> **Kantorovich Agentic Theorem:** A feasible setting of the process weights z is first order optimal (as defined previously) if and only if there exists a "shadow price" $\lambda_i \geq 0$ for each internal commodity i such that under the commodity pricing $p + \lambda$ each active process is operating at break-even (zero profit) and no inactive process is operating at a strictly positive profit. The shadow prices λ_i are Kantorovich's "objectively determined valuations" [Kan65] — imputed prices for internal commodities supplied by the optimization itself rather than by any human consumption market.

5 Summary

A total automation economy run by AI should serve humanity by providing what people want as indicated by the purchases that they make. Kantorovich showed that a certain notion of centralized optimization of industrial production (setting prices to zero for internal commodities) is essentially equivalent to free market economics with agents pursuing profit. This raises various questions. First, does this explain the value of factoring AI systems into agents — agentic AI — where each agent is following a different (profit?) motive. Second, do these economic principles apply to ecology and evolution where the system consists of organisms each with a "profit motive" of reproduction. Perhaps this also applies to systems of chemical catalysts each defining a production reaction.

The analysis presented here also indicates an alignment vulnerability under which the economy might encounter run-away development of internal commodities never subjected to human pricing. Presumably evolution and ecology are examples of systems of internal commodities with shadow prices unguided by any external system of market pricing.

Acknowledgments

I would like to thank Sergiy Verstyuk and Claude AI for useful comments on this paper.

References

[AD54] Kenneth J. Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica*, 22(3):265–290, 1954.

[Kan39] Leonid V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422, 1939. Russian original, Leningrad State University. English translation by R. W. Campbell and W. H. Marlow in *Management Science*, vol. 6, no. 4 (July 1960), pp. 366–422.

[Kan65] Leonid V. Kantorovich. *The Best Use of Economic Resources*. Harvard University Press, Cambridge, MA, 1965. Translated from the 1959 Russian original; source of the "objectively determined valuations" (shadow prices).

[Kar39] William Karush. Minima of functions of several variables with inequalities as side conditions. Master's thesis, Department of Mathematics, University of Chicago, Chicago, IL, 1939. M.Sc. dissertation; origin of the "K" in the KKT conditions.

[KT51] Harold W. Kuhn and Albert W. Tucker. Nonlinear programming. In Jerzy Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press, Berkeley, CA, 1951.