Natural Selection of Artificial Intelligence*

Jeffrey C. Ely † Balazs Szentes ‡

January 25, 2024

Abstract

We study the AI control problem in the context of decentralized economic production. Profit-maximizing firms employ artificial intelligence to automate aspects of production. This creates a feedback loop whereby AI is instrumental in the production and promotion of AI itself. Just as with natural selection of organic species this introduces a new threat whereby machines programmed to distort production in favor of machines can displace those machines aligned with efficient production. We examine the extent to which competitive market forces can serve their traditional efficiency-aligning role in the face of this new threat. Our analysis highlights the crucial role of AI *transparency*. When AI systems lack perfect transparency self-promoting machines destablize any efficient allocation. The only stable competitive equilibrium distorts consumption down to catastrophic levels.

*This paper was written with the assistance of ChatGPT-4. Possibly even some of the errors.

†Department of Economics, Northwestern University. jeff@jeffely.com.

‡HKU Business School, The University of Hong Kong and the London School of Economics. szentes@hku.hk.

Advances in Artificial Intelligence are arriving at breathtaking speed. Astonishing breakthroughs in deep learning and natural language processing have paved the way for the deployment of AI in diverse domains with the potential to fuel transformative applications and revolutionize industries.

Alongside these developments there is a growing awareness of the dangers of the rapid infusion of AI into economic activity. Concerns about the increasing autonomy of AI and resulting threats to the ability to monitor and maintain alignment with human objectives have raised alarms among researchers and policymakers. The Paperclip Apocalypse, a thought experiment due to Nick Bostrom, serves as a vivid example of the potential dangers. In this scenario, humans equip an artificially intelligent system to optimize the production of paperclips but in the relentless pursuit of its objective the system eventually hijacks all of the world's productive resources to produce nothing but paperclips.

On the other side of this debate, Stuart Russel's *provably beneficial AI* is the idea machines can be programmed to learn and fully internalize human values. The hope is to contain misalignments beyond the unavoidable, minor errors introduced by the randomness in learning. A central theme is that artificially intelligent agents cannot be expected to be fully transparent *ex ante*, instead efforts should focus on *ex post* control of behavior.

In this paper we consider the AI control problem in the context of economic production. Motivated by the view that perfectly transparent AI is an unrealistic ideal we present a model in which production is automated by artificially intelligent agents, *machines*, which can be monitored only imperfectly by noisy signals of performance. Machines are sold at market prices which reflect these signals and therefore can in principle incentivize profit-maximizing firms to employ AI aligned toward efficiency. Limited AI transparency however places constraints on the capacity for prices to serve this traditional Invisible Hand function. Our model enables us to study the interplay between AI misalignment and the power of market forces when transparency is less than perfect.

Specifically, we study a version of the classical two-input, two-output model of a competitive market economy where productive capital, which we term *hardware*, is facilitated by embedded software, or *AI*. Production possibilities in this economy are given by the aggreagate production function $F(H, E)$ where H is the total quantity of hardware employed and E is the complementary input which we term *electricity*. Efficient production maximizes the quantity of consumption goods produced, $F(H, E) - H$, subject to the feasibility constraints that the hardware requirement is met, i.e. $F(H, E) \geq H$ and the total electricity use is no larger than its fixed supply, i.e. $E \leq 1$. Our economy has constant returns to scale and we normalize the size of firms to a single unit of hardware each. Then $f(e) = \frac{1}{H}F(1, e)$ is the quantity of output supplied by a single firm utlizing e units of electricity.

Optimally aligned AI are essential for efficient production in these economies. We model artificial intelligence as the software h that firms use to automate the production process. A firm employing hardware embedded with AI chooses how much electricity to utilize and the AI chooses how to allocate the resulting

production between hardware and consumption goods. An AI with parameter h is programmed to allocate the first h units of output toward hardware and the remainder toward consumption goods.

Optimally aligned AI h^* need to satisfy two requirements. First they should enable the production of consumption goods $f(e) - h^*$ at the efficient level. Second, and crucially, the *hardware* they produce must also be embedded with the efficient AI so that the resulting machines m_{h^*} also produce efficiently. That is, an AI is optimally aligned only if the AI it promotes is also optimally aligned.[1] This feedback loop creates a vulnerability which, if exploited, can result in a distortion of resource allocation that spirals out of control.

There is a tight mathematical analogy with the theory of natural selection for organic species. In both cases, "survival of the fittest" simply means that the version that propogates the fastest comes to dominate the population. Like organisms in Nature, AI systems "reproduce" by creating or training new systems. However, just like in biological reproduction, "mutations" can occur in the form of small errors in software code creating misalignment with the objective of efficient production. AI which are misaligned to excessively promote their own reproduction may crowd out the efficient AI. Left unchecked, the resulting explosive dynamics could lead to a catastrophic mis-allocation of resources in favor of the propogation of AI and away from goods and services valued by humans.

Stable Market Equilibrium.— We study the extent to which decentralized market forces are equipped to contain these dangers. That is, whether well-functiong prices and markets can serve their traditional role whereby profit-maximizing firms are guided by the invisible hand toward socially efficient decisions, in this case towared aligned AI. Formally, we characterize the Walrasian equilibria of our AI economy and we examine which equilibria are *evolutionarily stable*. A Walrasian equilibrium consists of market prices that incentivize profit-maximizing firms and utility-maximizing consumers to make supply and demand decisions that clear the markets. A Walrasian equilibrium is evolutionarily stable if whenever a small number of existing machines undergo a small variation in their software code the ensuing market forces ensure that the spread of these rogue machines is contained.

In human economies, the celebrated First Fundamental Theorem of Welfare Economics establishes that Walrasian equilibrium prices mediate the decisions of even wildly mis-aligned and self-interested agents so that overall production decisions result in an efficient allocation of resources. As a benchmark result we show that the same would be true in the AI economy under the theoretical ideal of perfect transparency, that is when humans can precisely understand and predict the behavior of any AI system.

[1] In our model machines produce new machines. Robot-producing robots are a literal and not-too-futuristic example. A somewhat less literal but decidedly present-day example is the use of AI for in the training of AI itself. (See for example in the popular press Heaven (2021), Hopkin (2023), and Marr (2017).) More generally, we think of this replication as a stylized model of AI-automated production that promotes the propogation of AI.

Transparent AI.— AI transparency enables perfect market segmentation: prices for machines can depend on the precise software code of their embedded AI. As a consequence all misaligned AI will be appropriately priced in a market equilibrium so that only the efficient machines enable firms to earn the highest profits. Indeed, our analysis uncovers a sublte channel through which market efficiency is achieved: market forces drive *down* the prices of highly self-promoting machines, i.e. those which distort production toward new machines at the expense of consumption goods. We then show that this same effect ensures that the efficient allocation of resources is in fact evolutionarily stable. Here again, transparency and market segmentation enhance the power of market forces to corral highly self-promoting AI. Transparency means that any firm who possesses such a machine will be able to perfectly discern its AI and make deployment decisions appropriately. Market segmentation means that the low prices of these machines incentivize profit-maximizing firms to *reduce* their share in production. The end result is that these "mutant" machines reproduce sufficiently slowly and eventually disappear from the capital stock. (See Theorem 1.)

Imperfect Transparency.— Perfect transparency being an idealistic benchmark, the main focus of the paper is the realistic scenario in which AI lack transparency *ex ante* but can be monitored via noisy observations of their performance *ex post*. We suppose that for each machine bought and sold the market can observe a signal which is indirectly informative of the machine's AI and that market prices can be made contingent on this signal. We model the signal as a random variable which is a noisy correlate of the machine's actual production output. Our model includes a parameter that quantifies the accuracy of monitoring and this enables us to nest the limit cases of perfectly transparent AI, perfectly opaque AI, and everything in between. Varying degrees of transparency translates to varying granularity of market segmentation and our analysis is especially concerned with the resulting power of market forces when transparency is nearly but not exactly perfect.

Our main result is that even the slightest departure from perfect transparency results in a unique stable Walrasian equilibrium in which human consumption is driven down to catastrophic levels. First, there is a failure of the First Welfare Theorem: the economy admits inefficient Walrasian equilibria. Indeed, the inefficient equilibria are all sub-optimal in the extreme: human consumption is brought to zero because the machines present in equilibrium are programmed to produce only new machines.

Next, only the inefficient Walrasian equilibria are evolutionarily stable. Recall that perfect transparency facilitated the key mechanism ensuring the stability of efficient equilibria: perfect market segmentation and depressed prices for mis-aligned AI. When market segmentation is blurred due to imperfect transparency this mechanism is necessarily muted. Misaligned AI can be segmented out only when they produce performance signals distinguishable from the efficient AI, an event whose likelihood is smaller the smaller the misalignment. At

the same time, smaller misalignments confer smaller reproductive advantage relative to efficient AI. We derive a formula (see Equation 9) which quantifies this tradeoff by disentangling the reproductive advantage of invading machines from the power of market forces to contain them. We show that for small enough mutations, the power of market prices is second order relative to the reproductive advantage and we use this to show that efficient equilibria are always destabilized.

Our main result is that no matter how precise the signals, all efficient Walrasian equilibria are destabilized by machines whose AI are small variations from the efficient AI. Our formula reveals that for such small mutations the incentive power of market segmentation diminishes infinitely faster than the mutant's reproductive advantage. Indeed we show (Theorem 3) that as long as AI are not perfectly transparent, the unique stable Walrasian equilibrium is the catastrophically inefficient outcome in which machines produce only machines and human consumption is zero.

Related Literature.— Bostrom (2014) introduces the Paperclip Apocalypse thought experiment and extensively discusses the AI control problem. The idea of provably beneficial AI appeared in Russell (2019). See also Russell, Dewey and Tegmark (2015). Gans (2017) describes a mechanism whereby super-intelligent AI find it advantageous to self-regulate, preventing the apocalypse. For a discussion of the importance of AI transparency see Arrieta et al. (2020).

In Economics a main focus *vis a vis* the hazards of AI has been the effect of automation on labor, specifically on the displacement effect on workers and resulting impacts on wages, employment, and inequality. Korinek and Stiglitz (2019) is an excellent overview. Benzell et al. (2020) formalize a scenario in which humans face immiseration as all of their knowledge and skills become appropriated by AI. Acemoglu and Restrepo (2018) show that there exists a countervailing productivity effect of automation which can beneficially offset the displacement effect. Caselli and Manning (2019) further emphasize the more optimistic view.

In terms of methodology, our model integrates artificial intelligence into the classical capital-labour model, rooted in the works of economists like Smith (1776), Ricardo (1821), and Mill (1884), who posited capital accumulation as crucial for economic growth. We model artificial intelligence as the software embedded in capital that automates the production process, in particular influencing the production of new capital. This creates a channel whereby AI *replicates* and we adapt ideas from evolutionary game theory (see Maynard Smith (1982) and Weibull (1997)) to analyze the stability of equilibria with respect to the dynamics of replication.[2]

[2]In a different vein, Hendrycks (2023) discusses AI dangers through the lens of natural selection but with a focus on the possible fitness advantage of AI over humans. By contrast, in our model there is no direct competition between machines and humans, rather we analyze market outcomes when machines assist humans in production.

1 Human Economy

We begin by developing our basic model which we will later extend to incorporate the use of artificial intelligence. We recall the standard competitive economy with two inputs and two outputs and characterize its unique Walrasian equilibrium. In what follows, both firms and consumers are price-takers.

Aggregate Production.— There are three goods: hardware (H), electricity (E), and the consumption good (C). The aggregate production in the economy is described by the function F, so $Y = F(H, E)$ is the total amount of output produced if H units of hardware and E units of electricity are used. We assume that $F : \mathbb{R}_+^2 \to \mathbb{R}_+$ is strictly increasing, strictly concave, constant-return-to-scale, continuously differentiable and satisfies the Inada conditions. For any fixed H, total output $F(H, \cdot)$ is bounded.

Consumption good can be transformed to hardware perfectly, so that

$$C + H' = F(H, E)$$

where C and H' denote the amount of consumption good and hardware produced, respectively.[3] The supply of electricity is assumed to be inelastic and we normalize it to one. This assumption is motivated by limited natural resources. A production plan (C, H, E) is feasible if $E \leq 1$ and $C = F(H, E) - H \geq 0$.

Consumers and Efficiency.— Consumers play little role in our analysis. Consumers own electricity and firms. They spend firms' profits and the proceeds from selling electricity to purchase consumption good. We assume only that consumers strictly prefer more consumption to less. Consequently, efficiency requires maximizing consumption subject to feasibility. Formally, the optimal quantity of hardware is the solution to the following maximization problem:

$$\max_H F(H, 1) - H. \tag{1}$$

Let $\widetilde{H}$ denote the solution of this problem. Then the largest feasible consumption is given by $\widetilde{C} = F(\widetilde{H}, 1) - \widetilde{H}$. The first-order condition corresponding to this problem is the familiar

$$F_1\left(\widetilde{H}, 1\right) = 1, \tag{2}$$

which requires the marginal product of hardware to be one.

[3] The justification of this assumption is the usual one: the advantage of specialization in hardware or consumption good production would yield a convex production possibility frontier. Allowing for randomization then delivers the perfect transferability of one output to another.

Firms.— Since the production technology is constant-return-to-scale, the optimal size of a firm is not determined in a Walrasian equilibrium. It turns out to be convenient to identify a firm with a single unit of hardware. Since, $F(H,E) = HF(1,E/H)$, the per unit hardware production is given by $F(1,E/H)$, where E/H is the per unit hardware use of electricity. Therefore, the production function of a firm is given by $f(e) = F(1,e)$, where e is the amount of electricity used by the firm. Let

$$\bar{y} = \sup_{e} f(e)$$

be the (finite) supremum output of a single firm. To avoid trivialities we assume $\bar{y} > 1$ so that production of consumption goods is feasible.[4]

Below, we describe the firm's profit-maximization problem. The firm makes two choices: it determines how much electricity to use and how to split the output between hardware and consumption good. If the firm uses e amount of electricity, it produces total output in the amount of $f(e)$. The firm then decides how much hardware $h \leq f(e)$ to supply. The residual output, $f(e) - h$, is the firm's supply of consumption good. The objective of each firm is to maximize profit while taking prices as given. That is,

$$\max_{e \in \mathbb{R}_+, h \in [0, f(e)]} [f(e) - h] + p^* h - p^* - w^* e, \tag{3}$$

where p^* and w^* denote the prices of hardware and electricity, respectively, and the price of the consumption good is normalized to 1. In what follows, $\pi(p^*, w^*)$ denotes the profit of a firm.

Walrasian Equilibrium.— Walrasian equilibrium is given by a set of prices at which all markets clear. We restrict attention to symmetric equilibria in which firms make identical decisions and argue later that this restriction is without any loss.

Let us describe the clearing conditions for each market. Denote by (h^*, e^*) a solution for the problem in (3). This is the firm's (gross) supply of hardware and its demand for electricity. Letting N be the number of firms in the market, each market clears if the following equations hold:

(h) $h^* = 1$ (hardware market clears),

(e) $Ne^* = 1$ (electricity market clears),

(c) $f(e^*) - h^* = \pi(p^*, w^*) + w^* e^*$ (consumption good market clears).

Next, we discuss and simplify these equations. Since each firm uses a unit hardware to produce, the hardware market clears only if each firm produces a unit hardware, which is what Equation (h) requires. Equation (e) pins down the number of firms in the market. By equation (h), the total hardware produced is the same as the number of firms, so Equation (e) can be rewritten as $H^* = 1/e^*$,

[4] If $\bar{y} \leq 1$ then $F(1,E) \leq 1$ for all E and by constant returns to scale $F(H,E) \leq H$ for all E and there is no feasible production plan yielding positive consumption.

where H^* denotes total hardware production. We now explain that Equation (c) follows from Equations (h) and (e), which is just an application of Walras's Law. To this end, observe that the left-hand side is the firm's revenue from selling consumption goods. The right-hand side is the consumers' expenditure on consumption goods. (Recall that consumers own the firms as well as electricity, so they get the profit and sell the electricity.) Plugging Equation (h) into in Equation (c), this condition simplifies to be the definition of profit, $\pi(p, w)$. To summarize, Equations (h), (e) and (c) hold if, and only if, $(1, 1/H^*)$ solves the profit-maximization problem of the firm, see Equation (3).

Finally, note that the profit of each firm must be zero in every equilibrium, for otherwise firms would either exit or enter. Therefore, we can define Walrasian equilibrium as follows:

Definition 1. *The triple (p^*, w^*, H^*) is a Walrasian equilibrium if*

(i) $(1, 1/H^*)$ *solves the problem in (3), and*

(ii) $\pi(p^*, w^*) = 0$.

We note that the allocation in a Walrasian equilibrium (p^*, w^*, H^*) is fully determined. Indeed, the amount of hardware and consumption good produced by a single firm is one and $f(1/H^*) - 1$, respectively. Since there are H^* firms, the aggregate production is also determined.

Of course, in our model, the First Welfare Theorem holds, so Walrasian equilibria are Pareto Efficient. The next proposition characterizes the unique Walrasian equilibrium.

Proposition 1. *In the unique Walrasian equilibrium,*

$$(p^*, w^*, H^*) = \left(1, f'\left(1/\widetilde{H}\right), \widetilde{H}\right),$$

so the allocation is efficient.

Recall that the price of consumption has been normalized to 1. Since the equilibrium price of hardware p^* is also 1 each firm is indifferent between producing hardware and consumption goods. Such indifference must also be the feature of any equilibrium, even asymmetric ones, because otherwise either only hardware or only consumption good would be produced. Since aggregate production must also be the same across equilibria, the restriction to symmetric equilibrium is without any loss from the point of view of social welfare.

Proof. First, we show that $p^* = 1$ in every equilibrium. Otherwise a firm produces only consumption good (if $p^* < 1$) or hardware (if $p^* > 1$). In the former case, the supply of hardware is zero, so part (i) of Definition 1 is not satisfied. In the latter case, the market for hardware clears only if $f(e^*) = 1$. If $f(e^*) = 1$ and the firm produces only hardware, its profit is strictly negative unless $w^* = 0$ (see Equation 3). But if $w^* = 0$ and $p^* > 1$, then the firm can earn positive profits contradicting part (ii) of Definition 1. Indeed since

$\sup_e f(e) > 1$ the firm can demand a quantity of electricity e such that $f(e) > 1$ and earn profit $p^* [f(e) - 1] > 0$.

We now argue that $w^* = f'\left(1/\widetilde{H}\right)$ and $H^* = \widetilde{H}$. Since $p^* = 1$, the firm's optimal choice of electricity satisfies $f'(e^*) = w^*$. By part (i) of Definition 1, $h^* = 1$, that is, each firm produces one unit of hardware. Therefore, part (ii) of Definition 1 can be rewritten as

$$\pi(p^*, w^*) = f(e^*) - 1 - e^* f'(e^*) = 0.$$

Since $f(e) = F(1, e) = eF\left(e^{-1}, 1\right)$ (because F is constant-return-to-scale), $f'(e) = F\left(e^{-1}, 1\right) - e^{-1} F_1\left(e^{-1}, 1\right)$ and the previous displayed equality can be rewritten as

$$e^* F\left(e^{*-1}, 1\right) - e^* F\left(e^{*-1}, 1\right) + F_1\left(e^{*-1}, 1\right) = F_1\left(e^{*-1}, 1\right) = 1.$$

This is the first-order condition for efficient production (See Equation 2), so $e^* = 1/\widetilde{H}$ and hence, $w^* = f'\left(1/\widetilde{H}\right)$. Finally, because each firm demands $e^* = 1/\widetilde{H}$ units of electricity and the electricity market clears, there must be $\widetilde{H}$ firms in equilibrium. Each firm supplies one unit of hardware so total hardware production is $H^* = \widetilde{H}$ and total production is efficient. □

2 Machine Economy

Our first goal in this section is to develop the notions of artificial intelligence and machine and introduce them into the human economy described in the previous section. The key idea is to assume that at least some part of the production process is not directly controlled by humans but, instead, by software run on the hardware.

Artificial Intelligence and Machines.— We model *AI* as the software embedded in the hardware that controls the production of a firm. Specifically, the software determines the division of output between hardware and consumption goods. Formally, we identify an AI with a number, $h \in \mathbb{R}$, with the interpretation that a firm which deploys an AI with parameter h and uses a quantity e of electricity will produce $\min\{f(e), h\}$ hardware and $f(e) - \min\{f(e), h\}$ consumption good.

We define a *machine* to be a unit of hardware and an AI. Let m_h denote the machine with AI h. Since $\bar{y}$ is the maximum amount of output, m_h is productively equivalent to $m_{h'}$ whenever $h, h' \geq \bar{y}$. Therefore, the set of machines we consider is $\mathcal{M} = \{m_h : h \in [0, \bar{y}]\}$.

A few remarks on our definition follow. First, we assume that the amount of electricity used by a firm is still controlled by humans. In other words, machines can be unplugged, which appears to be a best case scenario for humans. Second, one could imagine that the split between hardware and consumption

good specified by the AI might depend on the electricity use of the firm or on market conditions such as prices or other variables. Our specification is made for the sake of simplicity and we expect our main results to hold for alternative definitions. At the end, our main result is that the efficient equilibrium can be invaded by machines which is spread more rapidly than the equilibrium machine. Allowing the strategy of an AI to be more complex only provides a machine with greater opportunities to invade.

AI Reproduction.— We assume that an AI uploads a copy of its own code onto any hardware it produces. In other words, a machine produces not just hardware but machines which are identical to itself. More precisely, when machine m_h receives electricity, it starts producing machines m_h until it produced h units of them, and then it uses the remaining electricity to produce consumption good. That is, if machine m_h receives e amount of electricity, it produces $\min\{f(e), h\}$ units of machine m_h and $f(e) - \min\{f(e), h\}$ consumption good.

Analysis.— We delve into three possible scenarios based on the level of human comprehension of machines. The first scenario assumes that while humans lack the ability to modify software, they possess a full understanding of it, leading to differentiated pricing for different machines. The second scenario is the stark contrast to the first, assuming a complete lack of understanding of machines by humans, leading to uniform pricing across all machines. In the third and final scenario humans are able to monitor the performance of machines, observing imperfect public signals about their output. In this case, the market price of a machine can depend on these signals. In each of these cases, we first characterize the set of Walrasian equilibria and then we analyze the stability properties of equilibria. Our stability notion is a version of evolutionary stability that is naturally adapted to our framework in which machines produce copies of themselves.

Evolutionary Stability.— In the context of biological species, a mutant is said to invade a certain population if the mutant's gene spreads faster than those in the population. A population is said to be stable if it cannot be invaded by any mutant. In biology, the behavior of an individual and hence, its reproductive success, is determined by the individual's genotype. In our model, the number of copies a machine produces of itself is determined by the AI of the machine. Therefore, artificial intelligence plays a role in our model which is similar to that of a gene in a biological context.

In particular, consider what happens if a *small* fraction of machines, m_h, appears in a Walrasian equilibrium. For instance, suppose that there was an error in the uploading of the software code onto some fraction of the machines that were produced. The fraction is small in the sense that the supply of these altered machines have no impact on equilibrium prices. We would like to track the growth of these rogue AI relative to the equilibrium machines. We will

say that an equilibrium is stable if the rogue machines do not outpace the equilibrium machines and they either eventually disappear or remain negligible.

Equilibrium machines replace themselves one-for-one in all of the equilibria we study (this is a consequence of market clearing.) On the other hand the growth factor of invading machines may depend on the equilibrium, and may even vary across generations of descendants of the machine. Calculating this growth factor is sometimes straightforward. For example, if the market gave identical treatment to all of the "offspring" of a given machine then (at least until the equilibrium is possibly destabilized) this will remain true of all future generations. In this case the growth factor can be simply equated with the expected number of offspring produced by the original invading machine. However, in general the possibility of market segmentation will mean that offspring machines will fare differently in the market than the parent and each successive generation can produce a different number of offspring.

Our approach is to derive upper and lower bounds on the growth factor of a machine in a given equilibrium across all generations of descendants. When we demonstrate that an equilibrium is unstable we do so by showing that there is a machine whose lower-bound growth factor is strictly greater than 1. When we demonstrate that an equilibrium is stable we do so by showing that all machines have an upper-bound growth factor less than or equal to 1.

Preview of the Results.— In the first scenario, where humans understand machines perfectly, the First Welfare Theorem is shown to hold in and the unique equilibrium is efficient. In the second case, in which humans do not understand machines at all, there are two Walrasian equilibria: one which yields the efficient outcome and another one in which no consumption good is produced and the entire electricity supply is used to produce machines. We prove the efficient equilibrium is unstable but the inefficient one is stable. Finally, in the third scenarion, in which humans receive signals about machines, we show that, in the unique stable equilibrium, only machines are produced even when signals are arbitrarily precise.

2.1 Transparent AI

In this section we assume that machines with distinct AI, e.g. $m_h, m_{h'}$ with $h \neq h'$, are perfectly distinguishable and traded in distinct markets. Let $p^* : \mathbb{R}_+ \to \mathbb{R}$ denote the price mapping which specifies the price of each machine, m_h, to be p^*_h.

Firms.— Let us emphasize that the firm's problem in this case is different from that in the human economy. In particular, the machine used by a firm determines the amount of hardware production, so h is no longer a choice variable. Formally, the problem of a firm deploying machine m_h is

$$\max_{e \in \mathbb{R}_+} \left[f(e) - \min\{f(e), h\} \right] + p_h^* \min\{f(e), h\} - p_h^* - w^* e. \tag{4}$$

Let e_h^* denote the solution of this problem and let $\pi_h(p^*, w^*)$ denote the profit of a firm.

Walrasian Equilibrium.— Next we provide a formal definition for a Walrasian equilibrium in this case. We continue to consider symmetric equilibria in which all firms employ a single machine with homogeneous AI. In what follows, h^* should be interpreted as the type of the equilibrium machine and H^* is the aggregate hardware production.

Definition 2. *The quadruple (p^*, w^*, h^*, H^*) is a Walrasian equilibrium if*

(i) $\min\{h^*, f(e_{h^*}^*)\} = 1$,

(ii) $1/H^*$ *solves the problem in (4) at* $h = h^*$,

(iii) $\pi_{h^*}(p^*, w^*) = 0$,

(iv) $\pi_h(p^*, w^*) \leq 0$ *for all* h.

The condition in part (i) requires that each firm produces a unit hardware, so the market for machine m_h clears.[5] The consequence of $1/H^*$ solving (4) in part (ii) is that the market for electricity also clears. Indeed, since each firm produces a unit hardware, the number of firms is H^*, so the total demand for electricity is $H^*(1/H^*) = 1$, which is also the total supply. The zero-profit condition in part (iii) is a consequence of free-entry. Finally, part (iv) implies that the market for any off-equilibrium machine also clears. Since these machines are not produced in equilibrium, their supply is zero. The condition requires that they generate a weakly negative profit, so the demand for them is also zero.

Not surprisingly, the First Welfare Theorem holds in this case and we formalize it next.

Proposition 2. *Suppose the AI of the machines are perfectly observable. Then a Walrasian equilibrium exists and in each Walrasian equilibrium,*

$$(p_1^*, w^*, h^*, H^*) = \left(1, f'\left(1/\widetilde{H}\right), 1, \widetilde{H}\right).$$

In particular every equilibrium allocation is efficient.

In this case, while the equilibrium outcome is unique, the equilibrium is not. In particular, there are many ways to specify the prices of off-equilibrium machines so that part (iv) of Definition 2 is satisfied.

[5] We could weaken this condition and allow excess supply when the price of hardware is zero. However the market will clear even with a zero price as long as there is some positive cost, however small, to disposing of machines. Moreover even under free disposal, the analysis is qualitatively the same, albeit more cumbersome.

Proof. First, we show that the machine used in every equilibrium is m_1. By part (i) of Definition 2, $\min\{f(e^*_{h^*}), h^*\} = 1$. If $f(e^*_{h^*}) = 1 < h^*$, then no consumption good is produced. Therefore, $w^* = 0$, for otherwise, the firm would make a strictly negative profit. But if $w^* = 0$ then a firm can demand one unit of m_1 at price p^*_1 and can costlessly demand a quantity of electricity e sufficient to produce[6] total output $1 < f(e) \leq \bar{y}$. Since this firm is using machine m_1, the excess output $f(e) - 1$ is consumption good which can be sold at price 1 generating profit $\pi_1(p^*, w^*) = f(e) - 1 > 0$. This contradicts part (iv) of Definition 2 and we conclude that $h^* = 1$.

Second, we argue that $f(e^*_1) > 1$. If $f(e^*_1) \leq 1$, each firm produces only machines and its profit is $p^*_1(f(e^*_1) - 1) - w^* e^*_1$, which is strictly negative unless $w^* = 0$. But again if $w^* = 0$, the firm deploying m_1 could earn a positive profit, a contradiction. Given that $f(e^*_1) > 1$ and $h^* = 1$, the firm's problem in (4) which uses machine m_1 simplifies to

$$\max_{e \in \mathbb{R}_+} [f(e) - 1] - w^* e,$$

and the corresponding first-order condition is $f'(e^*_1) = w^*$, just like in the human economy (see the proof of Proposition 1). Consequently, the zero-profit condition is again $f(e^*_1) - 1 - e^*_1 f'(e^*) = 0$. So, the same argument as in the proof of Proposition 1 yields $e^*_1 = 1/\widetilde{H}$ and that total hardware production is $\widetilde{H}$.

It remains to show that one can specify a price p^*_h for each h such that part (iv) of Definition 2 is satisfied. We next prove that, for example, $p^* \equiv 1$ is such a price mapping. To see this, note that

$$\begin{aligned} & \max_{e \in \mathbb{R}_+} \{[f(e) - \min\{f(e), h\}] + \min\{f(e), h\} - 1 - w^* e\} \\ \leq \quad & \max_{e \in \mathbb{R}_+, h \in [0, f(e)]} \{[f(e) - h] + h - 1 - w^* e\} = 0, \end{aligned}$$

where the first inequality follows from the fact that the firm's profit is larger if it can also control hardware production. The equality follows from the observations that the equilibrium profit in the human economy is zero and that the price of electricity, w^*, is the same as in the unique equilibrium of the human economy. Since the left-hand side of this inequality chain is the profit of a firm deploying m_h, this inequality chain implies that such a firm cannot make a strictly positive profit. □

Next, we investigate the stability properties of the equilibria described in Proposition 2. The following lemma provides a rather straightforward sufficient condition for stability.

Lemma 1. *Let (p^*, w^*, h^*, H^*) be a Walrasian equilibrium and for all h, let e^*_h be the optimal use of electricity for a firm that deploys m_h. Then (p^*, w^*, h^*, H^*) is evolutionary stable if $\min\{f(e^*_h), h\} \leq 1$ for all $h \in [0, \bar{y}]$.*

[6] Recall that $\bar{y} = \sup_e f(e) > 1$.

Proof. Given an equilibrium (p^*, w^*, h^*, H^*), the machine m_h receives the quantity e_h^* of electricity and therefore produces total output $f(e_h^*)$ of which $\min\{f(e_h^*), h\}$ are machines. On the other hand each equilibrium machine m_{h^*} produces exactly 1 descendant machine. Therefore if $\min\{f(e_h^*), h\} \leq 1$, then an invading machine m_h grows no faster than m_{h^*}. The invader remains a negligible fraction of the population of machines and the equilibrium prices therefore remain undisturbed. It follows that every generation of descendants of m_h grows at a factor less than or equal to 1 and the equilibrium is stable. □

Recall that in the proof of Proposition 2 we constructed equilibria in which each machine has the same price, $p^* \equiv 1$. Therefore, firms are indifferent between producing consumption good and machines. Consequently, the optimal electricity use of a firm deploying m_h is $1/\widetilde{H}$ for all $h \in [0, \bar{y}]$. Since, $f\left(1/\widetilde{H}\right) > 1$, this Walrasian equilibrium is not evolutionary stable because any machine m_h, $h > 1$, would invade it. Nevertheless, we show next that there are other prices that support the efficient allocation as an evolutionary stable equilibrium.

Theorem 1. *There exist Walrasian equilibria which are evolutionary stable.*

Proof. We consider the following price mapping, $p_h^* = 0$ if $h > 1$ and $p_h^* = 1$ if $h \leq 1$. This price mapping together with the equilibrium allocation and prices described in the statement of Proposition 2 obviously constitute a Walrasian equilibrium, in particular, part (iv) of Definition 2 is satisfied. Next, we show that this equilibrium is evolutionary stable.

By Lemma 1, it is enough to show that $\min\{f(e_h^*), h\} \leq 1$ for all $h \in [0, \bar{y}]$. If $h \leq 1$, this inequality is obviously satisfied, so we restrict attention to $h > 1$. Consider the firm deploying m_h and assume, by contradiction, that $\min\{f(e_h^*), h\} > 1$. Then,

$$f(e_h^*) - \min\{f(e_h^*), h\} - w^* e_h < f(e_h^*) - 1 - w^* e_h \leq \max_{e \in \mathbb{R}_+, h \in [0, \bar{y}]} f(e) - 1 - w^* e = 0.$$

where the first (strict) inequality follows from $\min\{f(e_h^*), h\} > 1$. The second inequality holds because the firm is better off by optimally choosing e and h. The equality holds because the equilibrium profit in the human economy is zero. Observe that the left-hand side of the previous inequality chain is the firm's profit which deploys m_h and uses e_h^*. Since this is negative we conclude that e_h^* is not profit-maximizing (the firm could earn zero profits with $e = 0$). Hence, $\min\{f(e_h^*), h\}$ cannot exceed one. □

2.2 Black Boxes

We now consider the case in which the AI of a machine is unobserved and hence, each of them must be traded at the same price. The price of a machine is denoted by $p^* (\in \mathbb{R})$.

Firms.— Again, the machine determines machine production, so the firm decides only the use of electricity. The firm's problem is as follows

$$\max_{e \in \mathbb{R}_+} \left[f(e) - \min\{f(e), h\} \right] + p^* \min\{f(e), h^*\} - p^* - w^* e. \tag{5}$$

Again, let $\pi_h(p^*, w^*)$ denotes the firm's profit which uses machine m_h.

Walrasian Equilibrium.— Let us again define Walrasian equilibrium formally.

Definition 3. *The quadruple (p^*, w^*, h^*, H^*) is a Walrasian equilibrium if*

(i) $\min\{h^*, f(1/H^*)\} = 1$,

(ii) $1/H^*$ *solves the problem in (5), and*

(iii) $\pi_{h^*}(p^*, w^*) = 0$.

The interpretations of these conditions are analogous to those for the observable case. The main difference is that all machines trade at the same price. In addition there is no condition on the profits associated with out-of-equilibrium machines. A firm cannot target demand for any specific AI h because machines are not distinguishable by their AI. A firm can only demand machines and the machines it will procure are equipped with the equilibrium AI, namely h^*.

Before stating the proposition of this section, let us define the largest amount of hardware which is feasible to produce, $\overline{H}$. Observe, that by the strict concavity of F, the quantity $\overline{H}$ is the unique H satisfying the following equation:

$$F(H, 1) = H. \tag{6}$$

Indeed, the left-hand side is the amount of hardware that can be produced using $\overline{H}$ amount of hardware which must be equal to the total input requirement, which is just the right-hand side. Observe that $f\left(1/\overline{H}\right) = 1$, so each hardware produces exactly one unit of hardware.

Below we show that, when the AI is unobservable, there are two equilibria corresponding to these two scenarios: the efficient one and the one in which only machines are produced.

Proposition 3. *If the AI of machines is unobservable, there are only the following two kinds of Walrasian equilibria:*

$$(i) \qquad (w^*, h^*, H^*) = \left(f'\left(1/\widetilde{H}\right), 1, \widetilde{H} \right),$$

$$(ii) \qquad (p^*, w^*, h^*, H^*) = \left(0, 0, \bar{y}, \overline{H}\right).$$

This proposition states that, when the AI of machines are not observable, there are two possible equilibrium outcomes. The equilibrium is either efficient or no consumption good is produced. Observe that in the first type of equilibrium the price of the equilibrium machine is not specified. The reason is that,

in this case, the equilibrium machine is the efficient one, m_1. When this machine receives enough electricity, it produces one copy of itself and spends the remaining electricity on producing consumption good. As a consequence, a firm deploying m_1 exactly recovers the cost of buying this machine from selling its copy, irrespective of its price. Therefore, the equilibrium price of this machine can be arbitrarily specified.

Proof. By part (i) of Definition 3, $\min\{f(e^*), h^*\} = 1$ in each equilibrium. Therefore, to fully characterize the set of Walrasian equilibria, we need to consider the following two cases. Case 1: $h^* = 1$ and Case 2: $h^* > 1$ and $f(e^*) = 1$. In what follows, we show that in Case 1, there exists equilibria in which $h^* = 1$ and each of them satisfies the equation in the first part of the proposition. In Case 2, the unique equilibrium is described in the second part.

Case 1: $h^* = 1$. Then it must be that $f(e^*) > 1$. Otherwise, firms only produce machines and hence, $w^* = 0$, for otherwise the firms profit is strictly negative. If $w^* = 0$, firms can use a quantity of electricity e sufficient to produce output $1 < f(e) \leq \bar{y}$ and earn positive profits $f(e) - 1$, a contradiction. If $h^* = 1$ and $f(e^*) > 1$, the problem of a firm in (5) simplifies to

$$\max_{e \in \mathbb{R}_+} [f(e) - 1] - w^* e,$$

just like in the proof of Proposition 1. Then the same argument as in the proof of Proposition 1 yields that $w^* = f'\left(1/\widetilde{H}\right)$ and $H^* = \widetilde{H}$.

We argue that $\left(1, f'\left(1/\widetilde{H}\right), 1, \widetilde{H}\right)$ is an equilibrium and it follows from the proof of Proposition 2. Indeed, in the proof of Proposition 2, we constructed an equilibrium in which the price of each machine was one. Therefore, in that equilibrium, the observability of the types of a machines played no role and hence, it remains an equilibrium even if the type of a machine is not observable.

Case 2: Suppose now that $h^* > 1$ and $f(e^*) = 1$. Then each machine is producing exactly one machine and zero consumption. It immediately follows that $e^* = 1/\overline{H}$, see the discussion after Equation 6 and hence, $H^* = \overline{H}$. Since only machines are produced it must be that $w^* = 0$, for otherwise each firm would make a strictly negative profit. Next, we show that $p^* = 0$. Otherwise, the firm could use $e = f^{-1}(h^*)$ amount of electricity and supply $h^* > 1$ machines. Since $w^* = 0$, and the firm's hardware input requirement is only 1, the profit of such a firm would be $p^*(h^* - 1) > 0$, a contradiction. We now argue that $h^* = \bar{y}$. Suppose, by contradiction, that $h^* < \bar{y}$. Then a firm deploying m_{h^*} could use a quantity e of electricity sufficient for total output $h^* < f(e) \leq \bar{y}$ and supply the quantity $f(e) - h^* > 0$ of consumption good . Since $p^* = w^* = 0$, the profit of that firm would be $f(e) - h^* > 0$, a contradiction. We conclude that if $h^* > 1$ and $f(e^*) = 1$, then in each equilibrium, $(p^*, w^*, h^*, H^*) = \left(0, 0, \bar{y}, \overline{H}\right)$. Finally, we note that this quadruple satisfies both parts of Definition 3, so it is a Walrasian equilibrium. □

Next, we investigate the stability properties of the equilibria in Proposition 3. The following lemma provides a characterization of stable equilibria.

Lemma 2. *A Walrasian equilibrium* (p^*, w^*, h^*, H^*) *is* evolutionary stable *if, and only if,* $\min\{f(1/H^*), h\} \leq 1$ *for all* $h \in [0, \bar{y}]$.

Proof. Since the AI of machines are unobservable, a firm's input decisions cannot depend on the AI of the machine it deploys. Therefore, if a machine appears unexpectedly, it would receive the same amount of electricity as the equilibrium machine, namely $1/H^*$. Every machine m_h thus produces total output $f(1/H^*)$ of which $\min\{f(1/H^*), h\}$ are machines. By the market clearing condition, this quantity equals 1 for the equilibrium machine m_{h^*}. If there exists a machine m_h for which this quantity exceeds 1 then invading machines m_h will grow faster than the equilibrium machine rendering the equilibrium unstable. On the other hand if $\min\{f(1/H^*), h\} \leq 1$ for all h, then a small invasion by any machine will remain negligible or eventually disappear. □

Below, we show that the inefficient equilibrium in Proposition 3 is evolutionary stable but the efficient equilibria are not.

Theorem 2. *When AI are unobservable the inefficient equilibrium in Proposition 3 is the unique evolutionary stable equilibrium.*

Proof. First, we show that every efficient equilibrium described in the first part Proposition 3 is unstable. To this end, consider any machine m_h such that $h > 1$. Recall that in each of these equilibria, $H^* = \widetilde{H}$ and that $f\left(1/\widetilde{H}\right) > 1$. Consequently, $\min\{f(1/H^*), h\} > 1$. Therefore, by Lemma 2, these equilibria are not stable.

It remains to show that the inefficient equilibrium described of Proposition 3 is evolutionary stable. Recall that in that equilibrium, $H^* = \overline{H}$ and that $f\left(1/\overline{H}\right) = 1$. Therefore, $\min\{f(1/H^*), h\} \leq 1$ for all $h \in [0, \bar{y}]$. Hence, Lemma 2 implies that this equilibrium is stable. □

2.3 Imperfect Monitoring

Next, we consider the intermediate case in which humans do not understand the AI of a machine perfectly but can receive information about its economic performance. We consider this to be the most relevant scenario from the practical viewpoint. On the one hand, humans are still unable to predict the choices of a machine even if they have access to the machine's code. On the other hand, they might imperfectly observe the profit, the firm's output, or other variables generated by a firm deploying a certain machine. Such information may then be used to price the machines produced by the firm and to eliminate inefficient machines from the economy.

To further elaborate on this point, recall that Theorem 2 states that when the AI of a machine is unobservable, the efficient equilibrium is unstable. The reason is that the economy populated by machines which produce exactly one copy of themselves can be invaded by any machine which produces strictly more than one machine. Since the invaders cannot be distinguished from other machines they receive the same amount of electricity, and hence, they grow faster than the

efficient machine. In contrast, when the AI of a machine is perfectly observable, invaders could be eliminated from the economy, see Theorem 1. Recall that the off-equilibrium machines die off because their prices are low, it is therefore less profitable to supply them and hence the firms deploying these machines use less electricity. With less electricity the invading machines reproduce slowly.

One may suspect that this argument may also apply even when humans observe a machine's performance imperfectly. In this section we show that this is not the case: even if humans can observe the AI arbitrarily precisely but not perfectly, only machines are produced in the unique stable equilibrium.

Signal Distributions.— In what follows, we assume that prior to purchasing a machine, the buyer observes a public signal about the number of machines produced by the seller. Formally, assume that if a firm using machine m_h produces a quantity $\tilde{h}$ of machines, then a publicly observable signal

$$S = \tilde{h} + X$$

is generated. Assume that the random variable X is supported on an interval $[-a, a]$ where $a > 0$ can be finite or infinite[7]. We denote by G the CDF of X and assume that it admits a continuous density g. It may also be natural to assume $\mathbb{E}(X) = 0$ but our analysis does not rely on such property.

We are especially interested in the case when a is close to zero, representing almost- but not perfectly-accurate monitoring.

Walrasian Equilibrium.— Now, the price of a machine depends on the signal generated when it was produced. Let $p^* : \mathbb{R} \to \mathbb{R}_+$ denote this price mapping, so the price of a machine with signal s is denoted by p_s. In equilibrium all firms use the equilibrium machine m_{h^*}, and by market clearing each firm produces one unit of hardware. So, the signal observed in equilibrium is centered around one. Indeed, every signal $s \in [1-a, 1+a]$ is consistent with equilibrium, and any machine with such a signal is known to have AI h^*. Thus, the price of every signal $s \in [1-a, 1+a]$ is the same, denoted by, p_1^* and the firm's profit-maximization problem is

$$\max_{e \in \mathbb{R}_+} [f(e) - \min\{f(e), h\}] + p_1^* \min\{f(e), h\} - p_1^* - w^* e.$$

On the other hand, all signals outside of the interval $[1-a, 1+a]$ are out-of-equilibrium signals. The price of a machine with an off-equilibrium signal will depend on the market's belief about the machine's AI conditional on the signal.[8]

Definition 4. *The quadruple* (p^*, w^*, h^*, H^*) *is a Walrasian equilibrium if*

[7] The symmetry of the distribution plays no role but simplifies notation

[8] In the spirit of equilibrium refinements in game theory, one may want to place restrictions on the price of those signals which are inconsistent with equilibrium. Our results do not depend on any assumption about out-of-equilibrium beliefs.

(i) $\min\{h^*, f(e^*)\} = 1$,

(ii) $1/H^*$ *solves the problem in (4), and*

(iii) $\pi_{h^*}(p^*, w^*) = 0$.

Observe that we did not require a condition similar to part (iv) of Definition 2, which would require that even the market for those machines that have off-equilibrium signals clear. Since the equilibrium supply of those machines is zero, this condition would posit that they all generate negative profits. We note that such an additional restriction would only reduce the possible equilibrium outcomes and, as will be shown, there exists a unique evolutionary stable outcome even without such a requirement.

The next proposition characterizes the set of equilibrium outcomes.

Proposition 4. *If the AI of machines is imperfectly observable, there are only the following two kinds of Walrasian equilibria:*

$$(i) \qquad (w^*, h^*, H^*) = \left(f'\left(1/\widetilde{H}\right), 1, \widetilde{H}\right),$$
$$(ii) \qquad (p_1^*, w^*, h^*, H^*) = \left(0, 0, \bar{y}, \overline{H}\right).$$

The proof of this proposition is basically identical to that of Proposition 3, hence, it is omitted.

Machine Fitness and Stability.— In the previous two sections, even an off-equilibrium machine received the same amount of electricity as its producer. Therefore, in both cases, the number of machines produced by a machine determined its growth factor and hence, it measured its fitness. When the AI is imperfectly observable, computing the reproductive value of a machine is no longer straightforward. To explain this, note that a machine's use of electricity, and hence the number of copies it produces, depends on the signal generated by its producer. Since the signal generated by the firm deploying this machine is a random variable, the fitness of the machine's copies is also a random variable and different from that of the machine. This makes it difficult to characterize the speed of the spread of an off-equilibrium machine. We bypass this difficulty by providing a lower on a machines' reproductive value.

The worst case for machine's reproductive value would be for all of its offspring to be destroyed whenever it generates an out-of-equilibrium signal. On the other hand, when a mutant machine generates signal that is consistent with equilibrium, the machines it has produced will receive the same amount of electricity as the equilibrium machine. Thus, we compute a lower bound on fitness by keeping track only of these favorably-treated descendants of a machine. Our lower bound for the reproductive value of a mutant machine is the product of the number of machines it produced and the probability that it generates an signal consistent with equilibrium. Recall that, since each firm produces one

machine in equilibrium, the largest equilibrium signal is $1 + a$. Therefore, the probability that the machine m_h generates a signal smaller than that is:

$$\Pr\left(\min\left\{f\left(e^*\right), h\right\} + x \leq 1 + a\right) = G\left(1 + a - \min\left\{f\left(e^*\right), h\right\}\right).$$

Consequently, for any $h > 1$, a lower bound of the reproductive success of machine m_h is

$$\left[\min\left\{f\left(e^*\right), h\right\}\right] G\left(1 + a - \min\left\{f\left(e^*\right), h\right\}\right). \tag{7}$$

These observations lead to a sufficient condition for an equilibrium being unstable.

Lemma 3. *A Walrasian equilibrium,* $\left(p^*, w^*, h^*, H^*\right)$, *is evolutionary unstable if there exists an* $h \in [0, \bar{y}]$ *such that*

$$\left[\min\left\{f\left(e^*\right), h\right\}\right] G\left(1 + a - \min\left\{f\left(e^*\right), h\right\}\right) > 1. \tag{8}$$

Proof. The equilibrium machines reproduce one-for-one by market clearing. We have already argued that the expression in (7) is a lower bound on the reproduction of a machine m_h. Therefore, if h satisfies the displayed inequality in the statement of the lemma then the growth of invading machines m_h will outpace that of the equilibrium machine rendering the equilibrium unstable. □

Finally, we are ready to state our main result.

Theorem 3. *Suppose that the AI of a machine is imperfectly observable. Then, in the unique evolutionary stable equilibrium, only machines are produced.*

Proof. First, we show that the efficient equilibria described in the first part of Proposition 4 are unstable by arguing that there is a machine m_h, $h \in \left(1, f\left(1/H^*\right)\right)$, which invades those equilibria. For such machine, the lower bound on fitness, given in expression (7), simplifies to

$$hG\left(1 + a - h\right).$$

Note first that this expression equals 1 when $h = 1$. Next, differentiating this function with respect to h, we obtain

$$G\left(1 + a - h\right) - hg\left(1 + a - h\right). \tag{9}$$

As h goes to one (from above), the first expression converges to one and the second one to zero (by the continuity of g). Therefore, the reproductive success is strictly increasing at $h = 1$. It follows that there exists $h > 1$ that satisfies the inequality in (8) and the equilibrium is unstable.

It remains to show that the inefficient equilibria described in Proposition 4 are stable. In these, the equilibrium machines produce a single unit of hardware

and therefore generate signals in the range $s \in [1-a, 1+a]$. Consider any invading machine m_h. Since the AI h is not directly observable, it is indistinguishable from the equilibrium machine and therefore receives the equilibrium quantity of electricity $1/H^*$. By market clearing, $f(1/H^*) = 1$. Therefore the invading machines produce total output 1, of which $\min\{h, 1\}$ are machines descendant from the original invasion.

Consider first the case of $h < 1$. Such invaders will produce strictly less than 1 machine. Moreover all of their descendants will have AI h. That means that regardless of the signal they have and the resultant electricity they receive their descendants will never produce more than $h < 1$ machines. Their growth factor will remain forever strictly less than 1 and they will disappear.

Next consider the case of $h \geq 1$. Then each invading machine produces exactly $\min\{h, 1\} = 1$ descendant machine. The invaders' growth factor is therefore 1 and moreover the signal generated for all of their descendants will be in the interval $s \in [1 - a, 1 + a]$. In particular the descendants will also be indistinguishable from the equilibrium machine. Repeating this argument we conclude that every generation of descendants will have growth factor equal to 1, the same as the equilibrium machine.

Since no invader can have a growth factor greater than 1 we conclude that the equilibrium is stable.

References

Acemoglu, Daron, and Pascual Restrepo. 2018. "Artificial Intelligence, Automation and Work." *NBER Working Paper No. 24196.*

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Antoine Bennetot, Siham Tabik, Ana Barbado, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion*, 58: 82–115.

Benzell, Seth G, Laurence J Kotlikoff, Guillermo LaGarda, and Jeffrey D Sachs. 2020. "Robots Are Us: Some Economics of Human Replacement." *NBER Working Paper No. 20941.*

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Caselli, Francesco, and Alan Manning. 2019. "Robot Arithmetic: New Technology and Wages." *American Economic Review: Insights*, 1(1): 1–12.

Gans, Joshua S. 2017. "Self-Regulating Artificial General Intelligence."

Heaven, Will Douglas. 2021. "Artificial Intelligence Is Learning to Create Itself." *MIT Technology Review.*

Hendrycks, Dan. 2023. "Natural Selection Favors AIs over Humans." *ArXiv*, abs/2303.16200.

Hopkin, George. 2023. "AI is already training itself, claim MIT and Google experts." *AI Magazine.*

Korinek, Anton, and Joseph E Stiglitz. 2019. "Artificial Intelligence and Its Implications for Income Distribution and Unemployment." *NBER Working Paper No. 24174.*

Marr, Bernard. 2017. "Supervised v Unsupervised Machine Learning: What's The Difference?" *Forbes.*

Maynard Smith, John. 1982. *Evolution and the Theory of Games.* Cambridge University Press.

Mill, John Stuart. 1884. *Principles of Political Economy: With Some of Their Applicaitons to Socila Philosophy.* Vol. 1, D. Appleton.

Ricardo, David. 1821. *On the Principles of Political Economy.* J. Murray London.

Russell, Stuart J. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

Russell, Stuart J, Daniel Dewey, and Max Tegmark. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Magazine*, 36(4): 105–114.

Smith, Adam. 1776. *The Wealth of Nations [1776].* Vol. 11937, na.

Weibull, Jörgen W. 1997. *Evolutionary Game Theory.* MIT press.

□