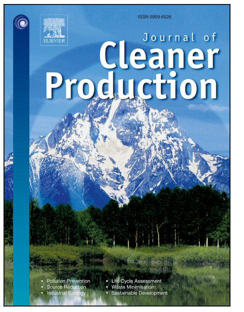
Accepted Manuscript



Growth, degrowth, and the challenge of artificial superintelligence

Salvador Pueyo

PII: S0959-6526(16)32201-6

DOI: 10.1016/j.jclepro.2016.12.138

Reference: JCLP 8713

To appear in: *Journal of Cleaner Production*

Received Date: 15 January 2016

Revised Date: 22 November 2016

Accepted Date: 23 December 2016

Number of words: 6,273 (6,073 without footnotes)

Growth, degrowth, and the challenge of artificial superintelligence

Salvador Pueyo[a,b]

[a] Department of Evolutionary Biology, Ecology, and Environmental Sciences, Universitat de Barcelona, Av. Diagonal 645, 08028 Barcelona, Catalonia (Spain)

[b] Research & Degrowth, C/ Trafalgar 8 3, 08010 Barcelona, Catalonia (Spain)

E-mail address: spueyo@ic3.cat

ABSTRACT

The implications of technological innovation for sustainability are becoming increasingly complex with information technology moving machines from being mere tools for production or objects of consumption to playing a role in economic decision making. This emerging role will acquire overwhelming importance if, as a growing body of literature suggests, artificial intelligence is underway to outperform human intelligence in most of its dimensions, thus becoming *superintelligence*. Hitherto, the risks posed by this technology have been framed as a technical rather than a political challenge. With the help of a thought experiment, this paper explores the environmental and social implications of superintelligence emerging in an economy shaped by neoliberal policies. It is argued that such policies exacerbate the risk of extremely adverse impacts. The experiment also serves to highlight some serious flaws in the pursuit of economic efficiency and growth *per se*, and suggests that the challenge of superintelligence cannot be separated from the other major environmental and social challenges, demanding a fundamental transformation along the lines of degrowth. Crucially, with machines outperforming them in their functions, there is little reason to expect economic elites to be exempt from the threats that superintelligence would pose in a neoliberal context, which opens a door to overcoming vested interests that stand in the way of social change toward sustainability and equity.

Keywords: Artificial intelligence; Singularity; Limits to growth; Ecological economics; Evolutionary economics; Futures studies

1. Introduction

We could be approaching a technological breakthrough with unparalleled impact on the lives of every reader of this paper, and on the whole biosphere. It might seem fanciful to suggest that, in a near future, artificial intelligence (AI) could vastly outperform human intelligence in most or all of its dimensions, thus becoming *superintelligence*. However, in the last few years, this position has been endorsed by a number of recognized scholars and key actors of the AI industry. Several research institutions have been created to explore the implications of superintelligence, for example at Oxford and Cambridge Universities. For details on how this idea emerged and is becoming established, see the chronological table in the Supplementary Material, and for a thorough understanding of the current discussions see Bostrom (2014) or Shanahan (2015).

Artificial intelligence (AI) is defined as *computational procedures for automated sensing, learning, reasoning, and decision making* (AAAI, 2009, p. 1). AIs can be programmed to pursue some given goals. For example, AIs programmed to win chess matches have been defeating human world champions since 1997 (Bostrom, 2014). Current AIs have narrow scopes, while a hypothetical superintelligence would be more effective than humans in pursuing virtually every goal. AI experts surveyed in 2012/13 assigned a probability of 0.1 to crossing the threshold of human-level intelligence by 2022, 0.5 by 2040 and 0.9 by 2075 (median estimates; Müller et al., 2016). The European Commission recently launched the €1 billion Human Brain Project with the intent of simulating a complete human brain as early as 2023, but its chances of success have been questioned (Nature Editors, 2015), and superintelligence is thought to be more easily attainable by engineering it from first principles than by emulating brains (Bostrom, 2014).

Following Yudkowsky (2001), the current discussion on the implications of superintelligence (Bostrom, 2014; Shanahan, 2015) is framed around two possibilities: the first superintelligences to emerge will be either *hostile* or *friendly* (depending on their programmed goals). In most authors' views, these would result in either the worst or the best imaginable consequences for humanity, respectively[1]. Much subtler distinctions apply to weaker forms of AI, but it is argued that intermediate outcomes are less likely for an innovation as radical as superintelligence (Bostrom, 2014, p. 20).

Hostile superintelligence is imagined as a result of failure to specify and program the desired goals properly, or of instability in the programmed goals, or less frequently as the creation of some illicit group. Therefore, it is framed as a technical rather than a political challenge. Most of the research is focused on ways to align the goals of a hypothetical superintelligence with the goals of its programmer (Sotala and Yampolskiy, 2015), without questioning the economic and political system in which AI is being developed. Kurzweil (2005, p. 420) is explicit in that an *open free-market system* maximizes the likelihood of aligning AI with human interests, and is leading a confluence of major corporations to advance an agenda of radical techno-social transformation based on this and other allied technologies (Supplementary Material). The benefits imagined from

1 The techno-utopia of a world ruled by friendly superintelligence reveals extreme *technological enthusiasm* and *technocracy*, in Kerschner and Ehlers' (2016) terminology. Technocracy is also apparent in moves to avoid public implication in this issue (Supplementary Material).

friendly superintelligence find an economic expression in rates of growth at an order of magnitude above the traditional ones or more (Hanson, 2001, 2008; Bostrom, 2014).

This view is akin to that of some authors within sustainability science, who take seriously the environmental challenges posed by economic growth, technological innovation and the functioning of capitalist markets, but seek solutions based on these same elements. Opposed to this position is the idea of degrowth (D'Alissa et al., 2015). Degrowth advocates hold a diversity of views on technology (see the Introduction to this special issue), but agree that indefinite growth is not possible if measured in biophysical terms, and is not always desirable if measured as GDP, both for environmental and for social reasons. Also, they are critical of capitalist schemes: to foster a better life in a downsized economy, they would rather support redistribution, sharing, democracy and the promotion of non-materialistic and prosocial values.

The challenges of sustainability and of superintelligence are not independent. The changing fluxes of energy, matter, and information can be interpreted as different faces of a general acceleration[2]. More directly, it is argued below that superintelligence would deeply affect production technologies and also economic decisions, and could in turn be affected by the socioeconomic and ecological context in which it develops. Along the lines of Pueyo (2014, p. 3454), this paper presents an approach that integrates these topics. It employs insights from a variety of sources, such as ecological theory and several schools of economic theory.

The next section presents a thought experiment, in which superintelligence emerges after the technical aspects of goal alignment have been resolved, and this occurs specifically in a neoliberal scenario. Neoliberalism is a major force shaping current policies on a global level, which urges governments to assume as their main role the creation and support of capitalist markets, and to avoid interfering in their functioning (Mirowski, 2009). Neoliberal policies stand in sharp contrast to degrowth views: the first are largely rationalized as a way to enhance efficiency and production (Plehwe, 2009), and represent the maximum expression of capitalist values.

The thought experiment illustrates how superintelligence perfectly aligned with capitalist markets could have very undesirable consequences for humanity and the whole biosphere. It also suggests that there is little reason to expect that the wealthiest and most powerful people would be

2 The perception of general technological and social acceleration is shared by authors close to degrowth (Rosa and Scheuerman, 2009) and by those concerned with superintelligence. The latter often suggest that acceleration will culminate in a *singularity*, related to the emergence of this form of AI (Supplementary Material).

exempt from these consequences, which, as argued below, gives reason for hope. Section 3 raises the possibility of a broad social consensus to respond to this challenge along the lines of degrowth, thus tackling major technological, environmental, and social problems simultaneously. The uncertainty involved in these scenarios is vast, but, if a non-negligible probability is assigned to these two futures, little room is left for either complacency or resignation.

2. Thought experiment: Superintelligence in a neoliberal scenario

Neoliberalism is creating a very special breeding ground for superintelligence, because it strives to reduce the role of human agency in collective affairs. The neoliberal pioneer Friedrich Hayek argued that the *spontaneous order* of markets was preferable over conscious plans, because markets, he thought, have more capacity than humans to process information (Mirowski, 2009). Neoliberal policies are actively transferring decisions to markets (Mirowski, 2009), while firms' automated decision systems become an integral part of the market's information processing machinery (Davenport and Harris, 2005). Neoliberal globalization is locking governments in the role of mere players competing in the global market (Swank, 2016). Furthermore, automated governance is a foundational tenet of neoliberal ideology (Plehwe, 2009, p. 23).

In the neoliberal scenario, most technological development can be expected to take place either in the context of firms or in support of firms[3]. A number of institutionalist (Galbraith, 1985), post-Keynesian (Lavoie, 2014; and references therein) and evolutionary (Metcalfe, 2008) economists concur that, in capitalist markets, firms tend to maximize their growth rates (this principle is related but not identical to the neoclassical assumption that firms maximize profits; Lavoie, 2014). Growth maximization might be interpreted as expressing the goals of people in key positions, but, from an evolutionary perspective, it is thought to result from a mechanism akin to natural selection (Metcalfe, 2008). The first interpretation is insufficient if we accept that: (1) in big corporations, *the managerial bureaucracy is a coherent social-psychological system with motives and preferences of its own* (Gordon, 1968, p. 639; for an insider view, see Nace, 2005, pp. 1-10), (2) this system is becoming *techno-social-psychological* with the progressive incorporation of decision-making algorithms and the increasing opacity of such algorithms (Danaher, 2016), and (3) human mentality

3 E.g., EU's Human Brain Project *is committed to driving forward European industry* (HBP, n.d.).

and goals are partly shaped by firms themselves (Galbraith, 1985).

The type of AI best suited to participate in firms' decisions in this context is described in a recent review in *Science*: *AI researchers aim to construct a synthetic* homo economicus, *the mythical perfectly rational agent of neoclassical economics. We review progress toward creating this new species of machine*, machina economicus (Parkes and Wellman, 2015, p. 267; a more orthodox denomination would be *Machina oeconomica*).

Firm growth is thought to rely critically on retained earnings (Galbraith, 1985; Lavoie, 2014, p. 134-141). Therefore, economic selection can be generally expected to favor firms in which these are greater. The aggregate retained earnings[4] *RE* of all firms in an economy can be expressed as:

$$RE=F_{\mathbf{E}}(\mathbf{R},\mathbf{L},\mathbf{K})-\mathbf{w}\cdot\mathbf{L}-(\mathbf{i}+\boldsymbol{\delta})\cdot\mathbf{K}-g. \quad (1)$$

Bold symbols represent vectors (to indicate multidimensionality). F is an aggregate production function, relying on inputs of various types of natural resources **R**, labor **L** and capital **K** (including intelligent machines), and being affected by environmental factors[5] **E**; **w** are wages, **i** are returns to capital (dividends, interests) paid to households, $\boldsymbol{\delta}$ is depreciation and g are the net taxes paid to governments.

Increases in retained earnings face constraints, such as trade-offs among different parameters of Eq. 1. The present thought experiment explores the consequences of economic selection in a scenario in which two sets of constraints are nearly absent: sociopolitical constraints on market dynamics are averted by a neoliberal institutional setting, while technical constraints are overcome by asymptotically advanced technology (with extreme AI allowing for extreme technological development also in other fields). The environmental and the social implications are discussed in turn. Note that this scenario is not defined by some contingent choice of AIs' goals by their programmers: The goals of maximizing each firm's growth and retained earnings are assumed to emerge from the collective dynamics of large sets of entities subject to capitalistic rules of interaction and, therefore, to economic selection.

4 Here (like, e.g., in Lavoie, 2014), *retained earnings* are the part of earnings that the firm retains, i.e., a flow. Other sources use *retained earnings* to refer to the cumulative result of retaining earnings, i.e., a stock.

5 And also by technology and organization, but these are not introduced explicitly because they are assumed to affect every term of this equation. The inclusion of **R** and **E** and their multidimensionality rely on insights from ecological economics (e.g., Martinez-Alier, 2013).

2.1. Environment and resources

Extreme technology would allow maximizing F in Eq. 1 for some given **R** and **E**, but would also alter the availability of resources **R** and the environment **E** indirectly. Would there still be relevant limits to growth? How would these transformations affect welfare?

To address the first question, let us consider growth in different dimensions:

- Energetic throughput: It is often thought that the source that could allow *energy production* (meaning tapping of exergy) to keep on increasing in the long term is nuclear fusion. This will depend on whether it is physically possible for controlled nuclear fusion to reach an energy return on energy investment EROI >> 1 (Hall, 2009). Even in this case, new limits would be eventually met, such as global warming due to the dissipated heat by-product (Berg et al., 2015). This same limit applies to other sources, such as space-based solar power. It is not known how global warming and other components of **E** would affect F in a superintelligent economy, or the potential for mitigation or adaptation with a bearable energetic cost. Whatever the sources of energy eventually used, the constraints on growth are likely to become less stringent right after the development of superintelligence, but this bonus could be exhausted soon if there is a substantial acceleration of growth.
- Other components of biophysical throughput: Economies use a variety of resources with different functions, subject to their own limits. However, extreme technological knowledge would allow collapsing the various resource constraints into a single energetic constraint, so energy could become a common numeraire. The mineral resources that have been dispersed into the environment can be recovered at an energetic cost (Bardi, 2010). Currently, many constraints on biological resources cannot be overcome by spending energy (e.g., the overexploitation of some given species), but this will change if future developments in nanotechnology, genetic engineering or other technologies are used to obtain goods reproducing the properties that create market demand for such resources.
- Information processing: Information processing has a cost in terms of resources. Operating energy needs pose an obstacle to brain emulations with current computers (Sandberg, 2016), but the hardware requirements (Sandberg, 2016) could be met soon (Hsu, 2016), and other paths to superintelligence could be more efficient (Sandberg, 2016). However, current ICT relies on a variety of elements that are increasingly scarce (Ragnarsdóttir, 2008). In principle, closing their

cycles once they are dispersed in the environment has an enormous energetic cost (Bardi, 2010). The resource needs of future intelligent devices are unknown, but could limit their proliferation. This does not have to be incompatible with a continued increase in their capabilities: When ecosystems reach their own environmental limits, biological production stagnates or declines, but, often, there is a succession of species with increasing capacity to process information (Margalef, 1980).

- GDP: Potentially, it could continue to increase without need of growth in biophysical throughput, e.g., through trade in online services. It is argued in Sec. 2.2 that this could well happen without benefiting human welfare.

Superintelligence holds the potential for extreme ecoefficiency: In the terms of Eq. 1, firms could not only increase F given $\mathbf{R}$, but also decrease depreciation $\boldsymbol{\delta}$ (which, however, would only be viable for assets that do not need quick innovation because of competition). Increasing resource efficiency and decreasing turnover are common in maturing ecosystems (Margalef, 1980). However, ecoefficiency does not suffice to prevent impacts on the environment $\mathbf{E}$ (which does not only affect production but also the welfare of humans and other sentient beings). With firms maximizing their growth with few legal constraints (as corresponds to the type of society envisaged in Sec. 2.2), extreme resource efficiency could well entail an extreme rebound effect (Alcott, 2014), which is tantamount to generalized ecological disruption.

2.2. Society

The literature on superintelligence foresees enormous benefits if superintelligent devices are aligned with market interests, including tremendous profits for the owners of capital (Hanson, 2001, 2008; Bostrom, 2014). By simple extrapolation of shorter-term prognoses (Frey and Osborne, 2013; see also van Est and Kool, 2015), this literature also anticipates huge technological unemployment, but Bostrom (2014, p. 162) claims that, with an astronomic GDP, the trickle down of even minute amounts in relative terms would result in fortunes in absolute terms. However, if there were astronomic growth (e.g., focused on the virtual sphere) while food or other essential goods remained subject to environmental constraints and competition between basic needs and other uses, resulting in mounting prices, a minute income in relative terms would be minute in its practical usefulness, and most people might not benefit from this growth, or even survive (think, e.g., of the role of biofuels in recent famines; Eide, 2009). In fact, there are even more basic aspects of the standard view that are debatable. This section presents a different view, building on the assumption that firms generally tend to maximize growth under environmental constraints. The following points discuss the resulting changes in each of the social parameters in Eq. 1, and relate them to broader changes in society:

- **L**: A continuing trend toward **L**=**0** is plausible, but it could also be reversed because of resource scarcity. Following Sec. 2.1, energetic cost could be the main factor to decide between humans or machines in functions that do not need large physical or mental capacities. Humans are made up of elements that follow relatively closed cycles and are easily available, while most current machines use nonrenewable materials whose availability is declining irreversibly (Georgescu-Roegen, 1971). Intelligent devices could thus become quite costly (Sec. 2.1). A variety of responses are imaginable, from finding techniques to build machines with more sustainable materials to creating machine-biological hybrids or modified humans; yet, it cannot be taken for granted that human work would be discarded. Initially, one extra reason to use human workers would be the big stock available. Even if human labor persisted, some major changes would be foreseeable: (1) Pervasive *rationalization* maximizing the output extracted from labor inputs. Current experience with digital firms point to insidious techniques of labor management to the detriment of workers' interests (Mosco, 2016). (2) AIs replacing humans in important functions that need large mental capacities. These include the senior managers of big corporations and other key decision makers (as well as people devoted to economically relevant creative or intellectual tasks). A few *unmanned* companies already exist (Cruz, 2014).

- **w**: Thus far, **w** and **L** seem to have been affected similarly by IT, via labor demand (Autor and Dorn, 2013). However, it is worth noting that firms also have an impact on human wants (Galbraith, 1985), and that this impact is being enhanced by AI. Every user of the Internet is already interacting daily with forerunners of *Machina oeconomica* that manage targeted advertising (Parkes and Wellman, 2015). *Relational artifacts* (Turkle, 2006) promise an even more sophisticated manipulation of human emotions. There is empirical evidence that, as it would be expected, the compulsion to consume induced by advertising results in longer working hours and depressed wages (Molinari and Turino, 2015). Furthermore, consumption is not the only motivation to work (Weber, 1904); e.g., some firms induce workers to identify with them (Galbraith, 1985). If these trends continued to the extreme, humanity would become extremely addicted to consumption and to work, and wages would drop to the minimum needed to survive and work (assuming that human labor remains competitive; otherwise, **w** would be reduced to the zero vector **0**).

- **i**: Like work, having capital invested in firms is not just motivated by the wish to consume (Weber, 1904). Procedures like inducing identification (Galbraith, 1985) could magnify the other motivations and reduce **i**. Consumption advertising acts in this case as a conflicting pressure (Molinari and Turino, 2015), but firms paying profits to households would probably be outcompeted by firms with no effective ownership (technically, nonprofits) or owned by other firms, which would allow reducing **i** to **0** (note that dividends and interests paid to other firms, including banks, cancel out because Eq. 1 refers to the aggregate of all firms). The owners of capital might currently have an economic function by allocating resources, but automated stock-trading systems have already determined between half and two thirds of U.S. equity trading in recent years (Karppi and Crawford, 2015), making human participation increasingly redundant.

- Demand: This is not an explicit term in Eq. 1, but is implicit in F to the extent that production is addressed to the market. In an economy in which humans receive minimum wages and no profits, or in an economy without humans, demand would be basically reduced to firms' investment demand. This would serve no purpose, but would result from economic selection favoring firms with the greatest growth rate. Given the complex interactions mediated by demand, it is unclear whether or not a maximization of each firm's growth should translate to a maximization of aggregate growth.

- g: For a strict neoliberal program, the main role of governments would be to serve markets, and this function would determine some g negotiated with firms. Directly or indirectly, governments would continue to exert functions of surveillance and coercion, aided by vast technological advances. Their decisions would be increasingly automated, whether or not they maintained some nominal power for human policy makers. Even elections are starting to be mediated by intelligent advertising (Mosco, 2016).

Therefore, a range of negative impacts can be expected, and they are unlikely to spare senior managers or capital owners.

Let us consider some moderate deviations from this political extreme. For example, these effectively “selfish” automated firms could coordinate to address shared problems such as resource limitations, but this does not mean that they would seek to benefit society, such as by ceding resources for people's use with no benefit for firms' growth. Or, before superintelligence is fully developed, governments could try to implement some model combining market competition as a force of technological innovation and wealth creation with economic and technological regulations to ensure that humans (in general, or some privileged groups) obtain some share of the wealth that is produced. However, this project would meet some formidable obstacles:

1. Ongoing neoliberal globalization is making it increasingly difficult to reverse the transfer of power to markets. A reversal will also be increasingly unlikely as computerization permeates and creates dependence in every sphere of life and the capacity of firms to shape human preferences increases.

2. The mere prohibition of some features in AIs[6] poses technical problems that could prove intractable. In the words of Russell (interviewed by Bohannon, 2015): *The regulation of nuclear weapons deals with objects and materials, whereas with AI it will be a bewildering variety of software that we cannot yet describe. I'm not aware of any large movement calling for regulation either inside or outside AI, because we don't know how to write such regulation*.

3. The objective role of humans obtaining profits from this type of firms would be parasitic. Parasites extract resources from organisms that surpass them in information and capacity of control (Margalef, 1980). In nature, parasites generally have high mortality rates, but persist by reproducing intensively. No equivalent strategy can be imagined in this case. The transfer of profits to humans would be an ecological anomaly, likely to be unstable in a competitive framework.

A much more likely departure from strict neoliberalism would result from structural mutations that would carry the system even further from any human plan, in unpredictable manners. Such mutations were excluded from the definition of this scenario, but not because they should be unlikely. In particular, they could provide a path to forms of *hostile superintelligence* more similar to those in the literature.

Marxists believe that societies dominated by one social class can be the breeding ground for newer hegemonic social classes. In this way bourgeois would have displaced aristocrats, and they expect proletarians to displace the bourgeois (Marx and Engels, 1888). However, the bourgeoisie represented an advance in information processing and control, unlike the proletariat. AIs are better

6 This would be one of the few types of regulation that appear to be acceptable from a neoliberal viewpoint, taking Hayek (1966) as a reference.

positioned to become hegemonic entities (even if unconsciously). This would not be just a social transition, but a biospheric transition comparable to the displacement of RNA by DNA as the main store of genetic information. So far, there is nothing locking future superintelligences in the service of human welfare (or the welfare of other sentient beings). Whether and how this future world would be shaped by the type of society from which it emerges is extremely uncertain, but neoliberalism can be seen as a blueprint for a Kafkaesque order in which humans are either absent or exploited for no purpose, and ecosystems deeply disturbed.

3. Degrowth as a viable alternative

Criticisms to the environmental and social impacts of the capitalist market are often answered with appeals to the gains in *efficiency* and long-term growth brought about by a *free* market. The above thought experiment shows how misleading it is to assume that efficiency and growth are intrinsically beneficial. The economic system as a whole may become larger and more efficient, but there is nothing in its *spontaneous order* guaranteeing that the whole will serve the interest of its human parts. This becomes even more evident when approaching the point in which humans could cease to be the most intelligent of the elements interacting in this complex system. Even though the thought experiment assumes neoliberal policies, as one of the purest expressions of pro-capitalist policies, Sec 2.2 also lists some reasons to be skeptical of reformist solutions.

Here, a response to this challenge is outlined. This involves, first of all, to disseminate it and integrate it into a general criticism of the logic of growth and a search for systemic alternatives, in contrast to the *technocratic* (*sensu* Kerschner and Ehlers, 2016) strategies to keep the management of this issue within limited circles (Supplementary Material). This awareness could initially permeate the social movements that originated in reaction to a variety of environmental and social problems caused by the current growth-oriented economy (including the incipient resistances to labor models introduced by digital firms; Mosco, 2016).

This will not just be one more addition to a list of dire warnings like resource exhaustion, environmental degradation and social injustice: While the economic elites now have the means to protect themselves from all of these threats, it is shown above that intelligent devices could well end up replacing them in their roles, thus equating their future to that of the rest of humanity. This alters the nature of the action for system change. It means that, in fact, this action does not oppose the interests of the most influential segments of society. A new role for social movements is to help these elites (and the rest of humanity) understand which policies are really in their best interest. In the kind of alternatives outlined below, such elites will gradually lose their privileges, but they will gain a much better life than if the loss of privileges occurs in the way that Sec. 2 suggests. Initially, few in the elites will be ready for such a radical change in their worldview, but these few could start a snowball effect. This is a game-changer creating new, previously unimaginable opportunities.

A key step will be to reform the process of international integration. Rather than democracy controlled by the market, markets will need to be democratically controlled (there has been a long-standing search for alternatives, e.g., The Group of Green Economists, 1992). This will not necessarily have to be followed by a trajectory toward a fully planned economy: a lot of research needs to be done on new ways to benefit from democratically *tamed* self-organization processes (Pueyo, 2014). What does not suffice, however, is the old recipe of setting some minimum constraints with the expectation that, then, the forces of market competition will be harnessed for the general interest. If, as suggested in Sec. 2.2, there is no way for governments to control a mass of entities evolving in undesirable ways, an alternative is to deflect the forces that drive such evolution. This entails nothing less than moving from an economic system that promotes self-interest, competitiveness, and unlimited material ambitions in firms and individuals to a system that promotes altruism, collective responsibility, and sufficiency. In short, moving from the logic of growth to the logic of degrowth (see D'Alissa et al., 2014).

Thus, besides regulations setting constraints of various types, there is a need for methods to align economic selection with the collective interests. The application of such methods would, for example, cause demand (which affects production F in Eq. 1) to become positively correlated with wages (i.e., with each firm's contribution to $\mathbf{w}$), negatively correlated with resource use ($\mathbf{R}$), and properly correlated with other more subtle parameters (not explicit in Eq. 1). The *common good economy* (Felber, 2015) is an approach worth considering because it aims explicitly to remove pressures that propel growth, and is already expanding with the involvement of many businesses. In this approach, a key tool is the *common good balance sheet*, a matrix of indicators of firms' social and environmental performance designed by participatory means, completed by the firms and (ideally) revised by independent auditors. Its function is to ease the application of ethical criteria by private and public agents interacting with firms in every stage of production and consumption. Felber (2015) envisions an advanced stage in which firms and the whole economy transcend their current nature (e.g., big firms would be democratized). While the common good balance sheet would serve mainly as an aid to change firms' general goals, it could also incorporate some explicit indicator of the perilousness of the software that these firms develop or use.

Hopefully, changing values in firms, governments, and social movements will also ease the change in individual values. This will further reduce the risk of having people engaged in the development of undesirable forms of AI. Furthermore, for those still engaged in such activities, there will be an increased chance of others in their social networks detecting and interfering with their endeavor. This reorientation at all levels (from the individual to the international sphere) will also help to address forms of AI distinct but no less problematic than *Machina oeconomica*, such as autonomous weapons.

Even with such transformations, it will not be easy to decide democratically the best level of development of AI, but the types of AI should become less challenging. (Also, these transformations could moderate the pace of technological change and make it more manageable, by relaxing the competitive pressure to innovate). However, they will only be viable if they take place before reaching a possible point of no return, which could occur well before superintelligence emerges (considering irreversibility, obstacle 1 in Sec. 2.2).

4. Conclusions

There is little predictability to the consequences that superintelligence will have if it does emerge. However, the thought experiment in Sec. 2 suggests some special reasons for concern if this technology is to arise from an economy forged by neoliberal principles. While this experiment draws a disturbing future both environmentally and socially, it also opens the door to a much better future, in which not only the challenges of superintelligence but many other environmental and social problems are addressed. This pinch of optimism has two foundations: 1) The thought experiment suggests that nobody is immune to this threat, including the economically powerful, which makes it less likely that the action to address it gets stranded on a conflict of interests. 2) The neutralization of this threat could need systemic change altering the very motivations of economic action, which would ally the solution of this problem with the solution of many other obstacles to a sustainable and fair society, along the lines of degrowth. One of the main dangers now lies in our hubris, which makes it so difficult to conceive of anything ever defying human hegemony.

Acknowledgements

I am grateful to Centre de Recerca Matemàtica (CRM) for its hospitality, to Melf-Hinrich Ehlers for calling my attention on Mirowski and other useful comments, to Aaron Vansintjan for proofreading the manuscript and for useful comments, and, also for their useful comments, to Linda Nierling , Laura Blanco, Anna Palau, Àlex Tortajada and the anonymous reviewers.

References

AAAI, 2009. Interim Report from the Panel Chairs. AAAI Presidential Panel on Long-Term AI Futures. Available at: https://www.aaai.org/Organization/Panel/panel-note.pdf (accessed 03-06-2015).

Alcott, B., 2014. Jevon's paradox (rebound effect), in: D'Alissa, G., Demaria, F., Kallis, G. (Eds.), 2015. Degrowth: A Vocabulary for a New Era. Routledge, London, pp. 121–124.

Autor, D. H., Dorn, D., 2013. The growth of low-skill service job and the polarization of the US labor market. Am. Econ. Rev. 103 , 1553-1597.

Bardi, U., 2010. Extracting minerals from seawater: an energy analysis. Sustainability 2, 980-992.

Berg, M., B. Hartley, Richters, O., 2015. A stock-flow consistent input–output model with applications to energy price shocks, interest rates, and heat emissions. New J. Phys. 17, 015011.

Bohannon, J., 2015. Fears of an AI pioneer. Science 349, 252.

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

Cruz, K., 2014. Exclusive interview with BitShares. Bitcoin Magazine, 8.10.2014. Available at: https://bitcoinmagazine.com/16972/exclusive-interview-bitshares/ (accessed 30.08.2015.).

D'Alissa, G., Demaria, F., Kallis, G. (Eds.), 2015. Degrowth: A Vocabulary for a New Era. Routledge, London.

Danaher, J., 2016. The threat of algocracy: Reality, resistance and accommodation. Philos. Technol., doi: 10.1007/s13347-015-0211-1.

Davenport, T.H., Harris, J.G., 2005. Automated decision making comes of age. MIT Sloan Manage. Rev. 46(4), 83-89.

Eide, A., 2009. The Right to Food and the Impact of Liquid Biofuels (Agrofuels). FAO, Rome.

Felber, C., 2015. Change Everything. Creating an Economy for the Common Good. Zed Books, London.

Frey, C.B., Osborne, M.A., 2013. The future of employment: How susceptible are jobs to computerisation? Oxford University. Available at: http://www.oxfordmartin.ox.ac.uk/publications/view/1314 (accessed 03.06.2015.).

Galbraith, J.K. 1985. The New Industrial State, 4th ed. Houghton Mifflin, Boston.

Georgescu-Roegen, N., 1971. The Entropy Law and the Economic Process. Harvard University Press, Cambridge, MA.

Gordon, S., 1968. The close of the Galbraithian system. J. Polit. Econ. 76, 635–644.

Hall, C.A.S., Balogh, S., Murphy, D.J.R., 2009. What is the minimum EROI that a sustainable society must have? Energies 2, 25–47.

Hanson, R.D., 2001. Economic growth given machine intelligence. Available at: http://hanson.gmu.edu/aigrow.pdf (accessed 09.08.2015.).

Hanson, R.D., 2008. Economics of the singularity. IEEE Spectrum 45(6), 45–50.

Hayek, F.A., 1966. The principles of a liberal social order. Il Politico 31, 601–618.

HBP, n.d. Overview. Available at: https://www.humanbrainproject.eu/2016-overview (accessed 28.04.2016.).

Hsu, J, 2016. Power problems threaten to strangle exascale computing. IEEE Spectrum, 08.01.2016. Available at: http://spectrum.ieee.org/computing/hardware/power-problems-threaten-to-strangle-exascale-computing (accessed 17.04.2016.).

Karppi, T., Crawford, K. 2016. Social media, financial algorithms and the Hack Crash. Theor. Cult. Soc. 33, 73–92.

Kerschner, C., Ehlers , M.-H., 2016. A framework of attitudes towards technology in theory and practice. Ecol. Econ. 126, 139–151.

Kurzweil, R., 2005. The Singularity Is Near: When Humans Transcend Biology. Duckworth, London.

Lavoie, M., 2014. Post-Keynesian Economics: New Foundations. Edward Elgar, Cheltenham, UK.

Margalef, R., 1980. La Biosfera entre la Termodinámica y el Juego. Omega, Barcelona.

Martinez-Alier, J., 2013. Ecological Economics, in: International Encyclopedia of the Social and Behavioral Sciences, Elsevier, Amsterdam, p. 91008.

Marx, K., Engels, F., 1888. Manifesto of the Communist Party (English version).

Metcalfe , J.S., 2008. Accounting for economic evolution: Fitness and the population method. J. Bioecon. 10, 23–49.

Mirowski, P., 2009. Postface, in: Mirowski, P., Plehwe, D. (Eds.), The Road from Mont Pèlerin. Harvard University Press, pp. 417–455.

Molinari, B., Turino, F., 2015. Advertising and aggregate consumption: A Bayesian DSGE assessment. Working Papers (Universidad Pablo de Olavide, Dept. Economics) 15.02. Available at: http://www.upo.es/econ/molinari/Doc/adv_rbc15.pdf.

Mosco, V., 2016. Marx in the cloud, in: Fuchs, C., Mosco, V. (Eds.), Marx in the Age of Digital Capitalism. Brill, Leiden, pp. 516–535.

Müller, V.C., Bostrom, N., 2016. Future progress in artificial intelligence: A survey of expert opinion, in: Müller, V.C. (Ed.), Fundamental Issues of Artificial Intelligence. Springer, Berlin, pp. 553-571.

Nace, T., 2005. Gangs of America. Berrett-Koehler, San Francisco, CA.

Nature Editors, 2015. Rethinking the brain. Nature 519, 389.

Parkes, D. C., Wellman, M. P., 2015. Economic reasoning and artificial intelligence. Science 349, 267-272.

Plehwe, D., 2009. Introduction, in: Mirowski, P., Plehwe, D. (Eds.), The Road from Mont Pèlerin. Harvard University Press, pp. 1–42.

Pueyo, S., 2014. Ecological econophysics for degrowth. Sustainability 6, 3431–3483. https://ecoecophys.files.wordpress.com/2015/03/pueyo-2014.pdf

Ragnarsdóttir, K.V., 2008. Rare metals getting rarer. Nat. Geosci. 1, 720–721.

Rosa, H., Scheuerman, W.E., 2009. High-Speed Society. Pennsylvania State University Press.

Sandberg, A., 2016. Energetics of the brain and AI. Tech. Rep. STR 2016-2. Available at: arXiv:1602.04019v1.

Shanahan, M., 2015. The Technological Singularity. MIT Press, Cambridge, MA.

Sotala, K., Yampolskiy, R.V., 2015. Responses to catastrophic AGI risk: A survey. Phys. Scripta 90, 018001.

Swank, D., 2016. Taxing choices: international competition, domestic institutions and the transformation of corporate tax policy. J. Eur. Public Policy 23, 571–603.

The Group of Green Economists, 1992. Ecological Economics: A Practical Programme for Global Reform. Zed Books, London.

Turkle, S., 2006. Artificial intelligence at 50: From building intelligence to nurturing sociabilities. Dartmouth Artifical Intelligence Conference, Hanover, NH, 15-07-2006. http://www.mit.edu/~sturkle/ai@50.html

van Est, R., Kool, L., 2015. Working on the Robot Society. Rathenau Instituut , The Hague.

Weber, M., 1904. Die protestantische Ethik und der “Geist” des Kapitalismus. Part 1. Archiv für Sozialwissenschaft und Sozialpolitik 20, 1-54.

Yudkowsky, E., 2001. Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. The Singularity Institute, San Francisco, CA. Available at: https://intelligence.org/files/CFAI.pdf