The Complex Economics of Artificial Intelligence

Juan Mateos-García (juan.mateos-garcia@nesta.org.uk)*

December 2, 2018

WORKING PAPER

Abstract

Artificial Intelligence (AI) systems help organizations manage complexity: they reduce the cost of predictions and hold the promise of more, better and faster decisions that enhance productivity and innovation. However, their deployment increases complexity at all levels of the economy, and with it, the risk of undesirable outcomes. *Organizationally*, uncertainty about how to adopt fallible AI systems could create AI divides between sectors and organizations. *Transactionally*, pervasive information asymmetries in AI markets could lead to unsafe, abusive and mediocre applications. *Societally*, individuals might opt for extreme levels of AI deployment in other sectors in exchange for lower prices and more convenience, creating disruption and inequality. *Temporally*, scientific, technological and market inertias could lock society into AI trajectories that are found to be inferior to alternative paths. New Sciences (and Policies) of the Artificial are needed to understand and manage the new economic complexities that AI brings, acknowledging that AI technologies are not neutral and can be steered in societally beneficial directions guided by the principles of *experimentation and evidence* to discover where and how to apply AI, *transparency and compliance* to remove information asymmetries and increase safety in AI markets, *social solidarity* to share the benefits and costs of AI deployment, and *diversity* in the AI trajectories that are explored and pursued and the perspectives that guide this process. This will involve an explicit elucidation of human and social goals and values, a mirror of the Turing test where different societies learn about themselves through their responses to the opportunities and challenges that powerful AI technologies pose.

*Nesta, 58 Victoria Embankment, EC4Y 0DS, London, United Kingdom. This essay has received valuable comments from John Davies and Chris Gorst, and benefited from conversations with Simone Vannucinni, W.E. Steinmueller and Tommaso Ciarli.

1 Introduction

We build powerful Artificial Intelligence (AI) systems to manage the complexity of our economies, and these systems make our economies more complex. This recursive loop holds important opportunities and risks that I explore in this essay.

Complexity creates demand for artificial intelligences when it surpasses the capacity of natural ones. Compare the pin factory in Adam Smith's *Wealth of Nations* with a modern, globally integrated and robotized industrial facility. The latter is much more complex: it involves many more actors, activities and interactions mediated by sophisticated technologies and responding to many more forces, such as changes in global demand, technologies and the behaviour of competitors. It generates more information that can be used to make more decisions and new types of decisions.[1] An obvious way to manage this complexity is by employing more workers to monitor and analyze the environment and act upon their insights. Humans are after all excellent at rapidly assessing and responding to new situations. But these individuals will be costly to hire and difficult to organize - imagine the army of workers and the level of organization that would be required to recommend products to users in an e-commerce platform such as Amazon. Abundance of information turns human attention into a scarce resource that has to be carefully managed [2]. Firms do this through routines and processes that act as organizational algorithms offering a menu of responses to different scenarios [2, 3].[2]

It is here that AI comes into play: Once trained on labelled data or artificial simulations, AI systems are able to partially mimic perceptive and flexible human decision-makers at a low cost, making predictions to inform action [4, 5].[3] This decrease in the cost of predictions increases their supply, and therefore the number of intelligent decisions that an organization can make economically, enabling personalization and interactivity in its products and services.[4] AI systems also help remove biases from human decision-making, and aid in the control of large technological systems such as scientific and industrial infrastructures and internet platforms generating amounts of data too vast and fast for human decision-makers. The potential applications are pervasive. This is why AI is being recognized as the latest example of a General Purpose Technology (GPT) with similar transformational potential to steam or electricity [7, 4].

And as was the case with those technologies, AI deployment brings with it dramatic changes that increase complexity at all levels of the economy.[5] It creates new interdependencies between the investments that organizations make and the practices they adopt, between the behaviours

[1] This is a manifestation of the Law of Requisite Variety (LRV), according to which a system needs a repertoire of responses as broad as the environment it seeks to manage [1].

[2] For example, the standard forms and procedures that help bureaucracies reduce diverse situations to a few cases.

[3] Complex environments also create more information that can be used to train AI systems in tasks with higher sample complexities - that is, tasks that are more diverse and where more data capturing a wider set of contingencies is required for effective performance [6].

[4] Note that this follows a functional definition of intelligence as adaptation to context rather than consciousness.

[5] I use the term deployment to refer to the development and diffusion of AI systems in organizations, markets and economies, together with complementary investments in skills, processes, business models and social and cultural attitudes and habits.

of actors in markets and groups in society, and between the decisions that are made over time. These interdependencies increase the risk of externalities, information asymmetries, coordination failures and strategic behaviours that could lead to undesirable outcomes from AI deployment. I review them in turn.

In the rest of this section I highlight related work, my contribution and normative standpoint, define AI and its link with Machine Learning (ML) and overview its potential impacts, highlighting why it is increasingly being recognized as the latest example of a transformative GPT. I also mention some features of AI that sets it apart from previous GPTs - in particular its fallibility (the fact that AI systems remain much narrower and more brittle than the human intelligences they try to emulate, and therefore liable to fail in unexpected situations) and its intangibility (which makes its deployment subtle and speedy).

Section 2 focuses on *organizational complexity*. AI requires complementary investments by organizations, institutions and individuals. For example, a firm has to select and implement an AI system, develop processes to turn its predictions into decisions and hire human workers with highly sought-after skills to manage this process and prevent errors - all of this while competitors watch closely to learn from its successes and avoid its mistakes. This is a process fraught with uncertainty that could hinder adoption, experimentation and knowledge sharing, particularly in those sectors where implementation is riskier (say, because it involves higher-stakes decisions) and/or more complicated because there are more organizations involved [8].

Section 3 shows that AI also increases *market complexity*. Uncertainty about AI impacts, low visibility in how it is being adopted and misalignment in incentives between actors create a thicket of informational asymmetries in AI markets - for example, AI researchers might opt for designing AI systems that perform well against existing benchmarks but are brittle and opaque in ways that limit their applicability. The firms that adopt AI systems know better than their users how personal data feeds their AI systems, and the metrics they are seeking to optimize. Malicious users will seek to manipulate the behaviour of AI systems for their benefit by feeding them spurious data. Behaviours like these could lead to races to the bottom in safety and respect for user rights, the abandonment of AI in some sectors, and the internalization of AI by organizations seeking to reduce transaction costs in 'AI lemon markets'.

Section 4 focuses on the *social complexity* brought about by information and power asymmetries between social groups. Individual choices in one area (such as buying a product based on AI) generates outcomes elsewhere (the conditions for workers in the sectors deploying AI systems). Individuals need to choose between cheaper, more convenient goods for them and disruption for other workers. Without coordination, they might opt for extreme AI deployment in the sectors where others work, and suffer it in their own. Different social groups will prefer others to bear the brunt of this disruption and seek to manipulate AI deployment to their advantage. This could increase inequality and create societal conflict.

In Section 5 I look at AI's *temporal complexity*: As AI systems and their complementary infrastructures are deployed, they create inertias in the research questions that are pursued, the techno-

logical architectures that are developed, the business models that are perceived as valid and even the human skills and capabilities that are valued. Some of these inertias will be hard to overcome in the future even if the trajectories chosen early-on, in a state of uncertainty and/or under the influence of opportunistic agents and vested interests, are found to be inferior to the paths not taken.

Together, all these factors create new economic complexities for human societies to understand (through research) and manage (through policy). I outline some principles for this research and policy effort in Section 6, under the rubric of *New Sciences (and Policies) of the Artificial* [2]. The foundational principle for this effort is that of *directionality*: AI technologies are not neutral. They can evolve in many different trajectories some of which are more desirable than others. Policy can play a role in steering AI deployment in a societally desirable trajectory based on the principles of *experiments and evidence* to measure AI impacts and identify their complementary investments; *transparency and compliance* to reduce the risk of unsafe and abusive outcomes in AI markets; *solidarity* to ensure that the benefits and costs of AI deployment are widely shared by different social groups and communities; and *diversity* in the scientific, technological and market AI trajectories that are explored so as to avoid premature lock-in to inferior paths of development.

Figure 1 summarizes key ideas in the essay.

1.1 Related work

I build on a growing body of literature on the economics of AI and in particular on the papers presented at two NBER workshops organized in 2017 and 2018 [9, 10]. Broadly speaking, those papers approach AI as a neutral and homogeneous General Purpose Technology (GPT) that greatly lowers the costs of prediction and will transform productivity and innovation once it is deployed together with suitable complementary investments [11]. The principal economic risk from AI deployment is that it will be too fast and disruptive, creating mass unemployment, rising inequality and political unrest [12].[6] With some exceptions, it is assumed that AI systems will eventually 'succeed' and that the main challenge will be for human societies to adapt to that success.[7]

By contrast to them, I place a stronger emphasis on situations where AI systems fail (at least to some degree) when they are deployed, creating costs that are unevenly distributed between social groups and over time. To do this, I draw on the literature on AI/Machine Learning (ML) safety and fairness risks [17, 16, 18, 19] and on key notions from evolutionary economics and Science, Technology and Innovation studies [20, 21, 22]. The AI risks literature has identified important failure modes in AI systems that could limit their generality, create hidden and unfairly distributed costs and require complementary investments (for example in supervision and monitoring) that should be taking into account during their economic analysis. Meanwhile, the evolutionary economics lit-

[6] Extreme scenarios where AI eliminates all jobs and brings the singularity are also considered in some cases [13, 14].

[7] The exceptions include [15], which considers the risk of mediocre AI systems that displace employment but do not result in sufficient improvements in productivity to augment demand for labour, [4], which acknowledges the potential for short term declines in user experience as organizations adopt AI systems, and [16], which approaches AI systems as agents whose behaviour might undermine the goals of the organizations that adopt them.

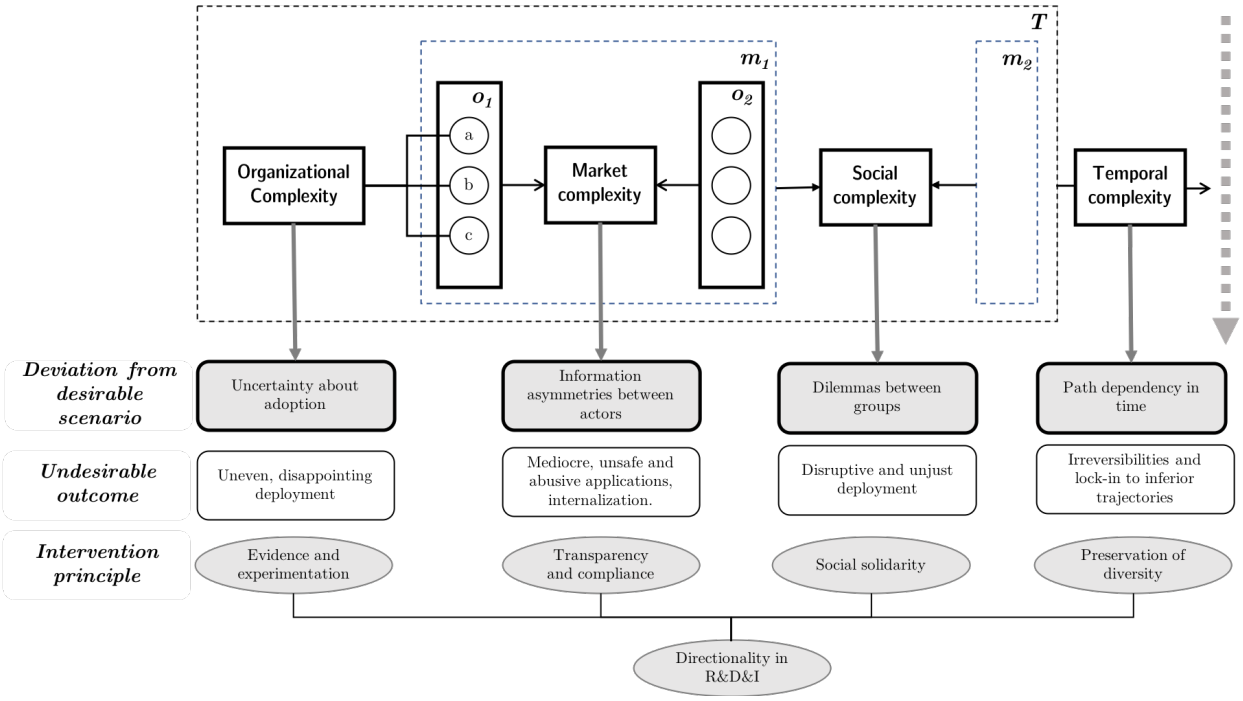

Figure 1: This figure summarizes key ideas in the essay: uncertainty about how to combine AI components (such as technologies, skills or processes) a, b and c in organization o_i creates organizational complexity. Information asymmetries about the nature of the inputs provided by other organizations in market m_i creates market complexity. Social dilemmas in the deployment of AI between sectors creates social complexity. Path dependence in decisions and investments in the formation of technological trajectory T creates temporal complexity. Each of these complexities creates the risk of deviations from an ideal scenario for AI deployment that can be addressed through policy interventions informed by the policy principles I set out in the essay.

erature suggests that technological trajectories are neither neutral nor homogeneous. They unfold in historical processes where accidents, mistakes and biases create path dependencies leading to potentially inefficient outcomes and externalities between generations [23]. These ideas lead me to pay more attention to the incentives and behaviours of AI researchers and scientists and to highlight the role of Research and Innovation (R&I) policy in AI deployment than is generally done in the literature.

Befitting the title of the essay, the resulting picture is more complicated than what one finds when considering the economic impacts of AI in silos or neglecting AI's fallibility and path dependence. Although my focus on economic failure modes for AI makes can make for somber reading in parts, this approach helps to identify what forces take us away from desirable, highly beneficial scenarios for AI deployment, as well as the research and policy principles that can help detect and remedy those deviations.

1.2 Normative standpoint: Rawlsian AI deployment

Throughout the essay, I refer to 'desirable' and 'undesirable' outcomes during AI deployment in organizations, markets, society and over time. These normative statements are based on the view that the deployment of AI should be just in a Rawlsian sense [24], in line with [13]. A just model for AI deployment would be accepted by a group of individuals making decisions behind a veil of ignorance where they did not know their position in society, their endowments and personal objectives. These individuals would accept that model for AI deployment if it enhanced (or at least did not reduce) their personal liberties and equality of opportunity, and if any increases in inequality that it brought were accompanied by improvements in the situation of the most disadvantaged members of society. This structure would strengthen broad-based social support for the deployment of AI systems with pervasive impacts.

At each level of deployment (organizations, markets, social groups, dynamic) I consider what features would be conducive to this outcome, and what forces take us away form it. The principles I propose in the conclusions support various (alternative or complementary) strategies to address those deviations, to be selected based on societal values and their practical effectiveness.[8]

1.3 AI definitions, applications and limits

1.3.1 What is it, and how does it work?

According to the Oxford Dictionary, Artificial Intelligence is the *'theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.'* This definition captures the goal of augmenting or automating different aspects of cognition that I alluded to in the introduction but tells us little about the shared characteristics of those tasks, or the nature of the technologies used to implement AI systems in practice. To make the discussion more general (in terms of the functions and tasks it covers) and specific (in bounding the components of AI systems), I define them as *'technological systems whose function is to inform or automate behaviours in dynamic situations.'* This definition captures the idea that AI systems need to be able to effectively adapt to a variety of situations - a hallmark of intelligence. Doing this requires *sensors* that collect data from the environment, *analyzers* that extract patterns from these data and recommend decisions that lead to changes in behaviour through *effectors*.

Computer and cognitive scientists have pursued different strategies to build AI systems since the 1956 *'Summer Research Project on Artificial Intelligence'* that arguably kick-started the field: Their initial approach (known as 'Good Old Fashioned AI') involved designing systems that process in-

[8] For example, AI disruption in labour markets could be alleviated by directing AI R&D towards labour-augmenting (rather than labour-displacing) applications, training those who have been displaced or compensating them through social policies. Public engagement and policy experimentation and learning between nations that follow different approaches could help identify a suitable policy mix leading to fair AI deployment. One obvious challenge I come back to in the conclusion is that this standpoint for AI deployment clashes with those that might be adopted in authoritarian societies, making it difficult to reach an agreement and coordinate development.

puts following sequences of 'if-then' rules hard-coded into the AI system [25]. In the 1980s, the 'expert system' approach sought to codify the decision-making processes of human experts into readily available knowledge management systems that would offer suitable responses to different situations [26]. Both approaches failed to deliver sufficiently flexible and reliable AI systems because the range of factors and exceptions to be taken into account in most complex decision-making scenarios (e.g. understand the context of a sentence to be able to translate between languages effectively) greatly exceeds the expressivity of most logical systems (e.g. the rules of grammar and dictionaries between languages). Further, many important cognitive faculties required for sensing the environment, such as object or speech recognition cannot be easily codified into discrete sets of rules. Paraphrasing Michael Polanyi, if we know more than we can tell, how will we impart that knowledge on AI systems?

In the last twenty years, Machine learning (ML) algorithms that learn patterns from training data (*supervised machine learning*) and develop successful strategies by trial and error (*reinforcement learning*)[27] have bypassed some of these limitations, leading to the development of more powerful AI systems. In these AI systems, assemblies of machine learning predictors act as sensors that use data to generate predictions about the environment. These predictions feed reward function analyzers that consider the benefits and cost of different scenarios: for example, given a credit card transaction that is identified as suspicious, what are the costs of ignoring it if it turns to be fraudulent versus the costs of inconveniencing a customer if it turns out not to be fraudulent?[9] These reward functions can be implemented into effectors that execute behaviours automatically (as is the case with a self-driving car or an AI targeted advert), or inform human workers who combine the AI prediction with their own judgment to reach a decision (like a radiologist who combines the AI analysis of a patient scan with her expert knowledge before recommending a treatment). See Figure 2 for a summary of AI systems' functions, capabilities and inputs.

The difference between AI and ML is subtle: AI systems often combine multiple ML algorithms to sense a range of dimensions of the environment and generate predictions that are integrated to inform decisions and actions [5]. An example of this would be a robot that uses computer vision, simultaneous localization and mapping and motion control to navigate an environment. It is also useful to think in terms of 'levels of AI'. We say that a system has a higher level of Artificial Intelligence when it requires less human input at key stages (i.e. there is a higher level of automation in decision-making and behaviour).

Recent advances in AI have been driven by increasing amounts of labelled data from internet sources, cheaper computation and storage from better hardware and cloud computing, and innovations in ML algorithms, particularly with the development of multi-layered neural networks that extract abstract patterns (features) from unstructured data such as images, video or sound with less need from human intervention (*deep learning*), and reinforcement learning systems that automatically adapt their strategies to feedback from the environment in order to optimize their performance [28].

[9] This example is taken from [4].

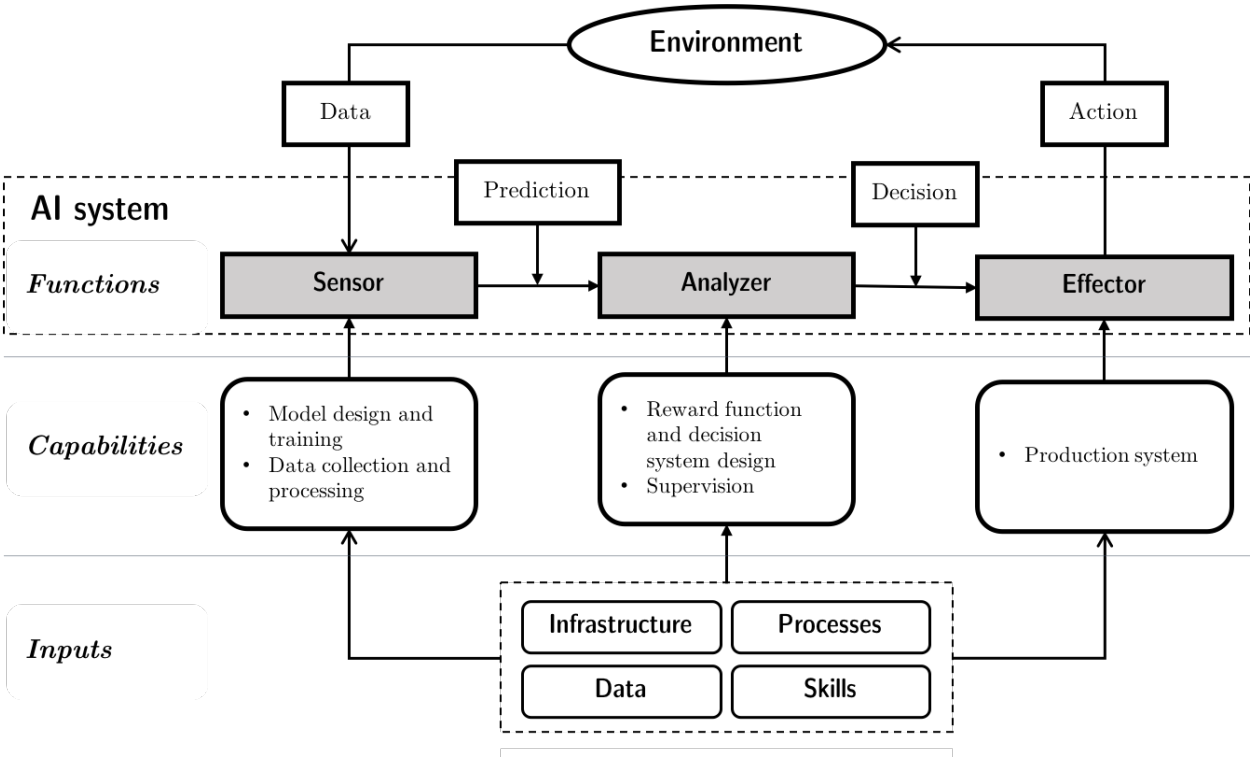

Figure 2: This figure represents the functions, capabilities and inputs of an AI system and its interactions with the environment.

All this has resulted in significant improvements in AI performance in domains such as computer vision, natural language processing, speech synthesis, mobility and game playing, and its application in many different industries [29]. Some examples include recommendation engines for retrieving and targeting information, images and video, classification systems that label images with comparable accuracy to human experts, natural language translation and recognition systems, and autonomous vehicles and robots that can operate in less controlled environments, and with less need for human supervision than was the case before.

In addition to imitating commonplace human capabilities at scale, AI systems can also go beyond human capabilities, thus augmenting human decision-making. For example, AI systems are, at least in theory, less prone to cognitive biases and over-fitting (inferring erroneous patterns from noisy data) than humans, so they can complement and enhance human decisions in domains such as recruitment, finance or the law [30, 31, 32].[10] AI systems can also parse big, fast data streams to control and optimize complex systems such as energy and digital infrastructures [33].

[10]This assumes unbiased input data and a stable decision environment. I consider the problems that appear when these conditions are not present next sub-section.

1.3.2 Where can it be applied?

AI's ability to automate and augment decision-making makes it widely applicable to many economic activities and industries. This is why it is being recognized as the latest example of a General Purpose Technology (GPT) like electricity or the combustion engine [11]: this means that it is technologically dynamic (susceptible of rapid improvements in performance), applicable in many industries, and capable of spawning follow-on innovations in these new application areas. Recent analyses of the development and diffusion of AI in different academic disciplines, computer science domains and industries supports this idea [7, 34]. The levels of R&D activity related to AI has increased, AI methods and techniques are diffusing into more sectors and disciplines, and they are proving influential wherever they do [10].

Ultimately, AI could greatly improve productivity by making the production and distribution of existing products and services more efficient, and powering new products based on automated decision-making. Amazon's AI-optimized supply chains and its virtual assistant Alexa illustrate both types of innovation. Even further, AI is an 'invention in the methods of invention' that could improve the productivity of scientific discovery and invention processes by, for example, enabling a faster and more comprehensive exploration of opportunities in research fields such as pharmaceutics or material science where innovation requires searching for valuable combinations in vast search spaces that can now be explored faster and deeper by AI systems [7, 35]. In doing this, AI could countervail stagnating productivity in science and technology (the fact that 'good ideas are getting harder to find') and drive productivity and economic growth in years to come [36]. The potential applications of AI are of course not limited to the private sector: AI could prove pivotal in tackling important social challenges from an aging population to environmental sustainability.

1.3.3 Where does it fail?

Although AI methodologies and ML algorithms are broadly applicable, once an AI system is implemented in a particular domain (that is, the system is trained on a particular dataset and integrated with a reward function for making decisions), it loses generality. This has several dimensions:

1. AI systems are *narrow*: AI systems need to be trained with large amounts of data from a domain. Their learning is highly specific to it, and difficult to transfer to other areas. The AI system for a self-driving car trained to operate in highways may not be suitable for cities.

2. AI systems are *brittle*: Related to the point above, the performance of AI systems declines if they are exposed to inputs that were not present in their training data. The AI system for a self-driving car trained in dry weather might break down when exposed to wet weather. There is a risk of AI errors and failures whenever an AI system is implemented in a real-world environment involving new situations, environments subject to change or environments where there are actors that want to manipulate the AI system [17, 37].[11]

[11] A salient case of this are *adversarial examples* that create catastrophic declines in performance when inputted into

3. AI systems are *greedy*: they maximize the amount of information they extract from the training data, but this creates the risk that they 'over-fit' and lose the ability to generalize to new information [39].[12] This also makes AI systems careless: they extract information from data regardless of its quality. An AI system trained on biased data will incorporate those biases into its model of the world. This can create problems when AI systems are applied in domains where existing data reflect social injustice or prejudices, as may be the case with the criminal justice system, university admissions, recruitment or access to credit [41, 42].[13]

4. AI systems are *mindless*: They literally optimize a pre-set performance metric or reward function even if this creates unexpected side-effects or contravenes the goals of their adopter [18, 16, 43]. This means that the goals of AI systems have to be carefully aligned with those of the organization deploying them in order to avoid surprising and unsafe AI behaviours.

5. AI systems are *opaque*: The optimization procedures that they follow to learn from data can be hard to interpret and explain. This makes it difficult to evaluate their safety, predict their behaviour and explain their outputs, rendering them unaccountable. Current theoretical understandings of the operation of state-of-the-art AI systems based on deep learning are still imperfect.[14]

1.3.4 How is AI different from other GPTs?

One feature of AI that sets it apart from previous GPTs is its relative intangibility: previous deployments of steam, electricity and information and communication technologies involved substantial investments in bespoke machinery and physical infrastructure. By contrast, key components of AI systems such as data, ML algorithms and reward functions are informational and therefore intangible.[15]

This intangibility leads to two new dynamics in AI deployment that underpin several situations and processes I will discuss in the rest of the essay: First, the adoption of AI can be done *subtly*: it is hard to determine, ex ante, whether a firm has implemented an AI system, to what degree, with what purpose and with what complementary investments: a robot based on sophisticated deep learning algorithms is hard to distinguish, initially, from one based on less advanced systems, and the recommendations generated by social networks using very different AI systems (and with

an AI system [38].

[12] ML researchers use methods such as *regularization* (penalizing excessively complex models) and *cross-validation* (evaluating models in test data-sets they were not originally trained on) to reduce overfit, and make their models more generalizable and robust [40].

[13] For example, if ethnic minorities are more likely to be arrested, or if they are more likely to re-offend due to discrimination in the labour market or policing, this will create a biased dataset. Model predictions will reflect these biases.

[14] This mirrors the situation in the early days of the Industrial Revolution, where practical (Ω) knowledge (know-how) about how to apply the steam engine raced ahead of theoretical (Λ) knowledge (know-why) about the physical processes underpinning its performance [44].

[15] This is not to say that AI systems do not have a physical substrate: they are stored and trained using increasingly specialized hardware, and some applications such as robots, self-driving cars and autonomous vehicles also require hardware effectors. But even in these cases, the intangibility of data and models means that some functions such as storage and processing can be rented flexibly from cloud computing services.

different goals) look quite similar in the surface. This makes it difficult to measure and monitor AI deployment and therefore estimate its impact, reduces the scope for spillovers (since it is hard for potential imitators to identify what combinations of practices are used by leading organizations, or even who those organizations are) and creates information asymmetries between different actors in AI markets. I discuss them in section 3.

Second, AI deployment can be done *speedily*: compared to other technologies requiring significant investments in infrastructure, it is relatively easy for an organization to adopt AI systems using widely available open source software implementations and train them with its own data or data available from web sources. This could lead to very fast rates of deployment that raise important risks if the deployment of AI creates hidden costs and side-effects, or irreversible outcomes. I study these situations in the rest of the essay.

2 Organizational complexity

TensorFlow, a popular Deep Learning software framework developed by Google engineers can be freely downloaded from GitHub, an open source software repository.[16] Does this mean that anyone can reap the economic benefits of AI simply by downloading, installing and starting to use this tool? Hardly. We know from the analysis of previous GPTs that successful deployment requires complementary investments in infrastructure, skills and processes - AI systems are *synergistic* [45]. What are some of their complements?

First, an AI system needs to be trained [28]. This requires labelled datasets or simulated environments about the domain where the AI system will be deployed, together with computational resources to store and process the data during training. AI systems are designed, implemented and tuned by experts with machine learning and software engineering skills. They are integrated with broader decision-making systems by decision-system designers, and their outputs are monitored for errors by supervisors [46, 4, 47]. Finally, the decisions they automate or inform need to be executed: this involves skilled workers who provide goods and services based on AI-informed decisions - for example, workers collaborating with robots in a warehouse.

Organizational processes, ways of working, structures and business models need to change too.[17] Previous studies of data-driven decision-making suggest that organizations with more and better data benefit from practices that empower employees to make decisions based on those data without consulting their supervisors [48, 49]. AI systems that easily redistribute information through an organization (for example, via portable devices) could strengthen the advantages of decentralization. However, brittle AI systems might require restrictions in worker agency and mobility to make their work environment more predictable, as is already done in robotized factories. Sophisticated supervisory and decision-making systems that keep humans in the loop help detect and

[16] https://github.com/tensorflow/tensorflow.

[17] In the same way in which impacts of electricity in industrial productivity did not materialize until factories radically reorganized their layout to reap the benefits of flexible electric motors, decades after its arrival [21].

remedy algorithmic errors, specially in high-stakes situations where mistakes have severe consequences [46].

In the rest of this section I show why the need to combine all of these technological, skills and organizational components makes the deployment of AI systems organizationally complex, and why this might lead to undesirable outcomes.

2.1 Desirable scenario

In the desirable scenario, organizations deploy AI systems in a way that secures their benefits while reducing their risks. This is possible because the suitability of an AI system for an organization's context and goals is well understood, as are the complementary investments that the organization needs to carry out in order to realize its benefits and manage its risks. Complementary skills are readily available in the market, and if new ones become necessary, educational institutions and trainers have the required information and incentives to address these changes in demand.

Organizations have incentives to carry out experiments with AI systems, and these experiments generate public knowledge that other organizations can learn from.

Suitable applications of AI appear in different industries and domains, both commercial and non-commercial, for the benefit of a broad range of consumers and constituencies. This parallel exploration of the potential uses of AI spawns further innovations that jump across sectors creating new applications, knowledge and benefits in a self-reinforcing process of innovation and learning.

2.2 Deviations

Here I summarize why organizational complexity could lead to deviations form the desirable scenario. Some of these ideas are formalized in the Mathematical Annex, and represented in Figure 3.

The first aspect of organizational complexity that leads to undesirable outcomes is that, as is the case with any innovation, there is uncertainty about the impacts of AI systems. While it is expected that the automation and augmentation of human decision-making could yield great benefits, this is based on estimates subject to error. This could lead to failures in deployment, particularly if there are systematic biases in actors' assessment of AI prospects, either because they are too optimistic (leading to failed deployment with disappointing benefits or unexpected accidents), or if they are too pessimistic (leading to aborted or excessively slow deployment that foregoes potential benefits from AI systems in some sectors). Differences in estimates of benefits and risks and risk-aversion across industries creates AI sectoral divides: AI systems are being deployed rapidly in technology and media sectors, and more slowly in sectors such as health or government where the commercial rewards are less immediate and the risks from errors are higher. Higher returns to AI deployment in fast adoption sectors could make AI components scarce in slow adoption sectors, further contributing to their divide.

Second, uncertainty about AI impacts is compounded by the need to deploy AI systems to-

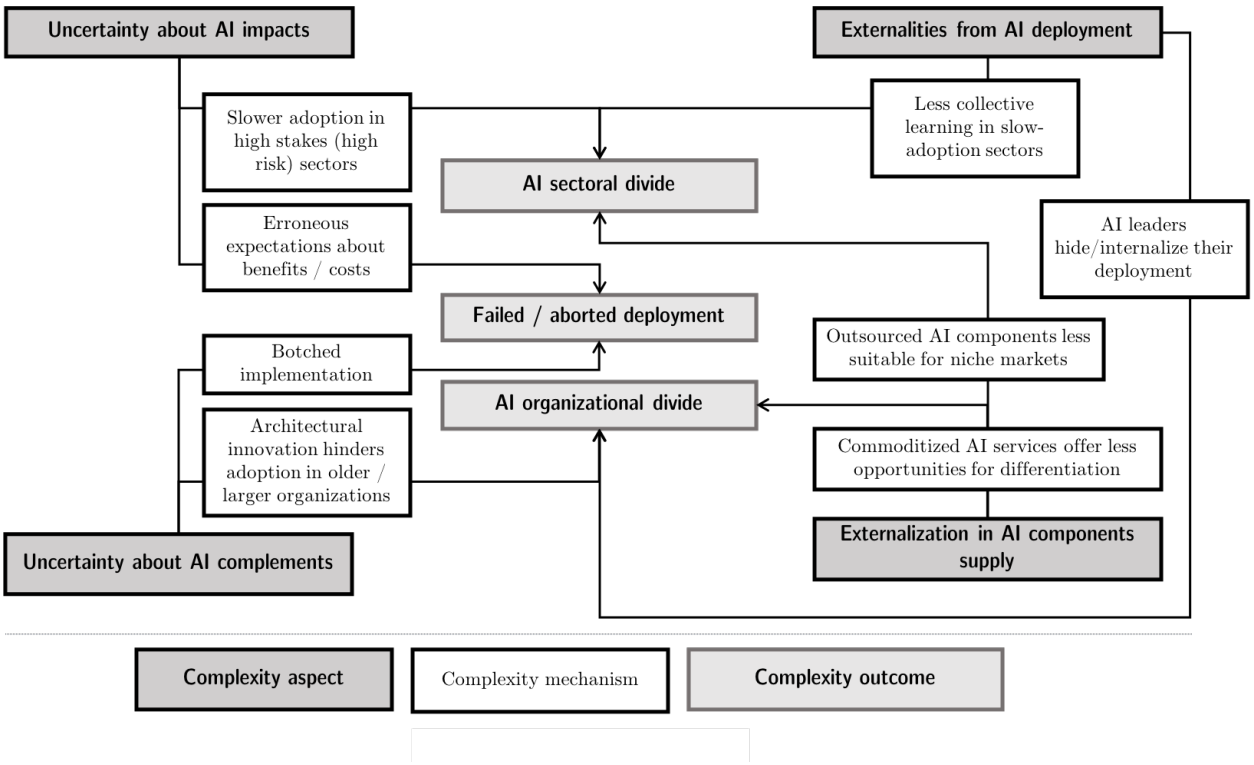

Figure 3: This figure presents the mechanisms through which different aspects of complexity impact generate undesirable outcomes

gether with investments in complementary technologies, skills and processes (I refer to these as AI complements), about which there is also uncertainty. The interactions between AI complements, and between AI complements and existing practices and processes increases the risk of disruption and failures in deployment. It could also result in an AI divide between those organizations that are more flexible and able to experiment, and those that are more rigid. A manifestation of this is the rapidity with which AI is being deployed in technology companies and start-ups, and the expanding productivity gap between them and other organizations. Another implication of uncertainty in AI complements is that those models for AI deployment that are 'simpler' and require fewer complements or less coordination with other actors (such as workers) could be preferred over other models that are more complex but perhaps also more beneficial.

Third, AI experimentation creates knowledge spillovers if organizations can free-ride on the experiments of innovators without suffering any of the risks. This could discourage experimentation, increase secrecy in deployment, or drive leading innovators that discover valuable AI components to secure control over them to restrict imitators or benefit from their entry (i.e. internalize the externalities from deployment). We see this in how leading adopters of AI systems in the technology sector are using their experience and know-how to build cloud computing infrastructures and platforms for AI deployment that they then sell to other organizations. Here, it is worth noting that subtlety and complexity in AI deployment are likely to hinder imitation by making it more difficult for followers to reverse-engineer the precise deployment model used by leaders. These

factors will further contribute to an AI organizational divide, and hinder AI deployment in slow adoption sectors lacking leaders to imitate.

Fourth, a market for AI components will develop. Suppliers that spread the costs of experimentation with AI components over larger markets can offer these components (and their combinations, such as cloud computing infrastructures that integrate data storage, processing and prediction) more cheaply and with less risk of failure. At the same time, the suitability of these outsourced components for an organization depends on its 'distance' from average market needs. Organizations with unique needs or in slow adoption sectors are less likely to be well-served by these external components, further dampening their AI deployment and increasing the risk of failures. This could also result in a 'dual market' for AI systems where leading organizations develop bespoke AI systems adapted to their needs, and offer commoditized systems with limited scope for innovation to mass market users. Once again, this increases the distance between leading adopters of AI and other organizations.

2.3 Complexity in deployment and skills

The interactions between AI deployment and skills have received much attention in the literature, where there is a general consensus that creative and social skills that complement AI deployment will be augmented by it, while routine skills are more likely to be displaced by it[50, 51, 52].[18] Displacement might not necessarily lead to a net loss of jobs if the associated improvements in productivity increase demand for other products from a sector, or elsewhere in the economy.[54, 15]

The organizational complexities that I have discussed in this section also interact with these labour market outcomes: first, firms may prefer to deploy labour-displacing AI systems over labour-augmenting alternatives if this requires less coordination with other actors such as workers or the suppliers of skills whose independent decisions will determine successful implementation.[19] Second, workers and educators also need to make decisions about what skills to learn / supply in a state of uncertainty about business decisions that will determine if these skills are augmented or displaced: they have to infer demand for skills from the complex process of organizational deployment described above [55]. Third, there is uncertainty about the actual impacts on productivity of the AI systems that are eventually adopted. It is possible that firms overestimate the benefits of AI systems and implement mediocre AI systems, or that insufficient and/or erroneous complementary investments generate mediocre impacts on productivity which prove insufficient to offset labor displacement [15].[20] Coordination in the supply of skills that complement AI systems becomes harder in this state of uncertainty, potentially hindering AI deployment and its benefits.

[18] This also has important implications for inequality through the creation of 'hourglass' shaped labour markets divided between highly productive and remunerated occupations that complement AI, and low productivity, low remuneration occupations that are difficult to automate.[53]

[19] Labour-displacing AI systems could be simpler to implement if they require less changes in interdependent labour practices - this will depend on the regulatory framework.

[20] Next section I consider the externality aspect of this.

3 Market complexity

How can a firm know what it is getting when it transacts with another in an AI market? AI subtlety makes it hard to determine what AI system has been adopted, for what purpose with what complementary practices and processes, and with what impacts. This creates market complexity manifested in a thicket of information failures leading to unsafe, mediocre or abusive AI applications, the abandonment of AI systems in some markets, or their internalization by a small number of powerful firms.

I use the term AI 'market' expansively, to refer to various activities related to the development, adoption and application of AI systems in market transactions, as well as their regulation. Some participants in these markets include:

- **AI agents** are the algorithms implemented in AI systems. Although they lack 'agency' in the traditional sense, they have design goals (metrics and rewards functions to optimize) and ideal goals (the goals of the organization implementing them, which inform the design goals).
- **Scientists** who research new and improved AI systems in public research organizations and universities, and in private sector laboratories and R&D units. AI scientists seek to advance the state of knowledge about AI systems and their performance, gain acclaim from their peers and attract research funding from the public and private sector.
- **Adopters** are organizations that deploy AI systems to achieve commercial, public or social goals. Here, I also include other important actors that shape the commercial application of AI systems, such as investors who fund AI ventures.
- **Individuals** who create data that is used by AI systems, interact with AI-derived predictions and decisions, and work in environments where AI has been or might be adopted. They have multiple, potentially conflicting goals, such as maintaining and improving their personal and political rights, working conditions and salaries, and accessing affordable, convenient and safe goods and services.
- **Government** refers to the public authorities of the country where the AI market is based.[21] It provides public services and funds public goods (such as basic research) and uses various instruments to encourage societally beneficial behaviours and discourage illegal behaviours.[22] It is represented in AI markets by public agencies such as research funders or regulators.

[21] AI intangibility means that many AI markets are international and therefore involves participation by multiple governments with different goals, further increasing the complexity of the situations I describe.

[22] I acknowledge this is a stylized, idealistic view of the role and behaviour of governments.

3.1 Desirable scenario

In the desirable scenario for AI deployment in markets, scientists develop AI systems using reproducible methods that others can review and assess, taking into account social and business needs and clearly communicating the strengths and limitations of the systems they develop.

AI systems are transparent and well understood, easy to implement in a way that is aligned with organizational goals, and unlikely to generate unexpected effects (or at least the types of algorithms and contexts where such effects are likely to occur are well understood, so that risks can be managed).

Adopters select those AI systems that are most valuable for them and implement them transparently and safely, together with suitable complementary investments (in line with the desirable scenario in 3). In doing this, they take into account the rights of their workers, as well as any wider impacts that the AI systems they adopt might have, including in labour markets.

Users and consumers are aware of how AI systems are being deployed in the organizations they interact or transact with, and the implications that this has for their personal data and the products and services they are offered. They are thus able to make informed decisions about what products and services to consume.

All of this takes place in a way that is consistent with societal goals and the regulations to uphold them.

3.2 Deviations

Information asymmetries take AI markets away from the desirable scenario: each transaction in an AI market is a principal-agent situation where the principal (for example, a government agency procuring an AI system to predict the risk of recidivism in the criminal justice system) has limited information about the menu of strategies available to the agent (in the previous case, the company that developed and trained the model). This creates incentives for the agent to follow a strategy that maximizes her benefits at the expense of the principal, or that creates costs and risks for the principal. The principal might itself be the agent in another transaction (e.g. the adopter agency deploys AI systems leading to unfair and discriminatory decisions against individuals). In this noisy market, it is difficult to determine who is liable for algorithmic errors if/when they happen.

The rest of this sub-section provides examples of information failures in transactions between actors in AI markets (see the Mathematical Annex for a formalization and 4 for a visual representation).

Manipulation and subversion of AI systems

Narrow and brittle AI systems can be manipulated and subverted by other actors. For example, these actors might want to use an AI service in a way that contravenes its terms of service, access valuable information from a secure AI platform, or degrade its performance for financial or politi-

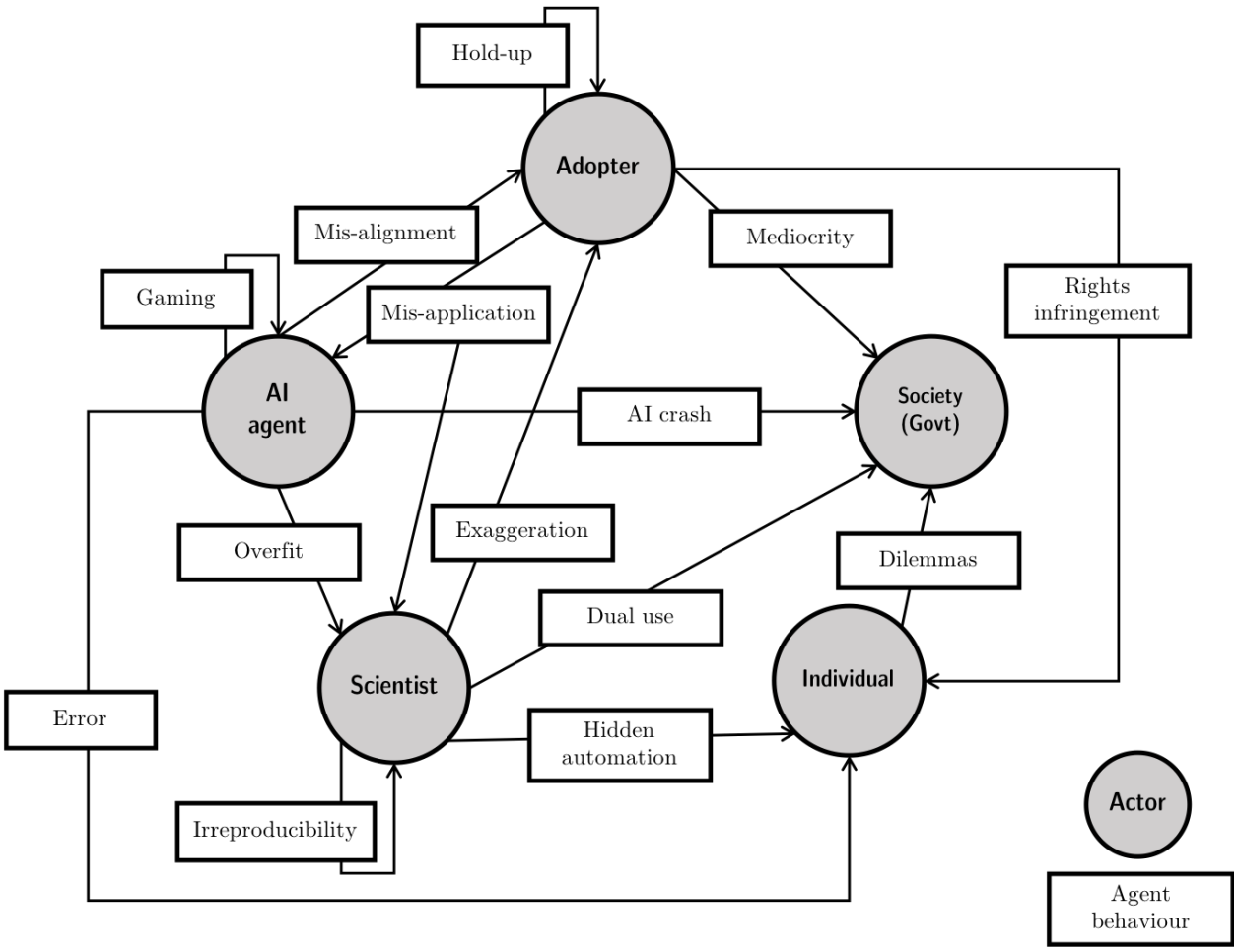

Figure 4: This figure presents information thickets between different actors in AI markets.

cal reasons. I use the term *Gaming* to refer to situations where an actor manipulates an AI system to behave against its ideal goals. For example, the actor might use misleading language to distribute spam or fake news in a social media platform, or implement a cyberattack using a strategy that mimics a regular pattern of user behaviour [56, 17]. The gaming actor might be an individual or an AI agent such as for example a bot or a Generative Adversarial Network designed to produce synthetic data hard to classify accurately by the AI defender [38]. *Sabotaging* refers to situations where individuals inside an organization (for example its workers) engage in acts that directly damage AIs or degrade their performance.

Risky deployments

Risky AI deployments are situations where insufficient information about the real performance and goals of an AI system lead to deployments with reduced benefits, hidden costs or insufficient safeguards. Although the organization adopting the AI system might not be seeking to benefit from its riskiness, it could still benefit economically if unsafe AI systems are cheaper to adopt than

safe ones.

Some risky implementations of AI includes cases where an AI agent *overfits* the data, extracting noisy patterns from it, and creating a divergence between expected and real performance leading to lower benefits and/or higher costs. *Unsafety* refers to interactions between AI systems and individuals that create risk for the latter [41]. A *Misaligned* AI agent pursues goals that diverge from those of its adopter. This could be because the AI reward function is insufficiently or imperfectly specified, or the AI has been trained on biased data, creating an 'omitted payoff bias' (optimizing the wrong metric) [30, 16, 18].

The final type of AI safety risk is a *AI Crash* where AIs operating in complex environments with many other actors, both human and algorithmic, behave in unexpected ways leading to system failures.[23] This is an example of a situation where the deployment of AI systems generates hidden costs that are not realized until they surpass a threshold value, and the crash is triggered [46].

Irresponsible Research and Innovation

In this case, an actor behaves irresponsibly or fraudulently in the market: she does this by exploiting, withholding or neglecting information about the characteristics and impacts of the AI systems that she is developing or adopting. She is able to do this because the costs of her behaviours are hidden or incurred by other actors in the transaction or interaction, or in the broader AI market and society [58].

One case is *Misapplication*. Here, an AI adopter deploys a narrow AI system in a task that is unsuitable for it because it involves new data it was not trained on, because the stakes are higher than in its original domain (the consequences of mistakes are more severe) or because it requires more interpretability or explainability in outputs than in the source domain. *Exaggeration* refers to situations where scientists overstate the benefits of an AI system or understate its costs / risks, potentially leading to its misapplication by adopters, or to an over-supply of funding for research to deliver these benefits.[24]

I use the term *Irreproducibility* to refer to a variety of behaviours that hinder scientific progress in AI [59]. This includes the dissemination of research papers without the materials required to reproduce the results so as to impede competing scientists or to hide flaws in the research. Also the use of ambiguous language or unnecessary complexity, techniques that over-fit to benchmark datasets, a preference for complex and opaque models with high performance in known metrics over more explainable and transparent approaches whose performance is harder to measure, or a bias towards publishing novel results. All these behaviours create uncertainty about the benefits, costs and limitations of new AI methods, and hinder learning between scientists.

Dual-use risks stem from the generality of AI, which creates the possibility that AI systems de-

[23] An example of this would be the 2010 'flash crash' in the New York Stock Exchange, where High-Frequency-Trading (HFT) algorithms began reacting to each other's actions in a high-speed recursive loop that crashed the market [57]

[24] This was the cause of previous 'AI winters', with drastic declines in funding for AI research after disappointments with its performance.

veloped for socially beneficial goals (e.g. computer vision systems for self-driving vehicles) will be applied in ethically questionable or dangerous ways (such as mass surveillance or autonomous weapons) [17]. This failure may take place between research funders and scientists who carelessly or opportunistically develop dangerous AI systems, or between scientists and malicious adopters who deploy AI systems in unexpected ways. Since much AI research is disseminated in public academic journals, and/or implemented (or implementable) in freely available open source software packages, this makes it harder to control how and where AI research outputs are applied [60]. *Hidden automation* is a situation where the actions of individual consumers and workers generate data which are, without their knowledge or permission, used to train AI systems that automate labour.

Finally, *Mediocrity* refers to situations where adopters implement AI systems with disappointing impacts . In addition to systems that are misapplied, unsafe or prone to generate AI crashes and unexpected failures, this also includes AI systems that degrade the quality of an user's experience while saving costs for their adopter, or mediocre AI systems that displace labour without substantial gains in productivity (this results in an externality for society in terms of increased welfare costs, societal conflict etc.) [12, 15, 13].

Exploitation and hold-up

Situations of exploitation occur when AI deployment infringes on the rights of individual users, consumers, or workers, or abuses a dominant position (in terms of market power or access to information) over a business partner.

This includes applications of AI that infringe on privacy, misuse personal data, attempt to manipulate the behaviours of individuals or encourage them to engage in addictive behaviours, use the data they supply in ways that they would perceive as unfair (such as for example through price discrimination of information discrimination), as well as applications that put individuals at risk (particularly when they are in positions of vulnerability) or degrade their working conditions.

On the business-to-business side, cases of hold-up include situations where an AI adopter trading with another organization manipulates prices, exploits its data or abuses a dominant position due to control of strategic assets such as data , IT infrastructure or access to consumers. It also includes situation where a provider of AI components exaggerates their benefits or their applicability for their client, resulting in their mis-application.

Dilemmas

Dilemmas refers to situations where the uncoordinated behaviour of individuals and social groups leads to societally undesirable outcomes. I focus on these situations next section.

3.3 Some outcomes

What could be the outcome of all these information thickets? Given the large number of actors involved, and the complex interdependences between their behaviours, we could envisage many possible situations. I sketch three of them.

First, there is the possibility of a *race to the bottom* with careless and/or unfair AI deployment through unsafe and/or exploitative practices and business models. There are few incentives to invest on making AI systems safe and protect user rights, and those organizations that do cannot compete with less principled rivals. Unsafe AI systems display unexpected behaviours when they are deployed in high-stake domains and complex social systems, resulting in AI crashes and increased systemic risks. Scientists develop powerful but potentially unsafe and unethical AI systems atop a knowledge base that is weakened by dubious research practices. In spite of this, individuals continue using and consuming AI goods and services because they are essential or addictive, or because they lack information about the way AI systems are implemented and their hidden costs and risks.

The second scenario is *abandonment*. In this case, individuals exit AI markets that are perceived to be unsafe and abusive, perhaps in response to a catastrophic event or mass AI failure. Funding for AI research stops or is curtailed, and the use of AI systems is restricted to a small number of applications where their performance and conditions for effective operation are well understood and strictly regulated.

A third scenario is *internalization*: the information failures above increase transaction costs in AI markets. Participants need to verify and validate any AI systems and AI-based products and services they procure, manage the risk of hold-up and algorithmic failure due to exploitative and unsafe implementations of AI systems by suppliers and clients, and reassure consumers against the perception of a race to the bottom in AI deployment. To reduce these transaction costs, these organizations internalize AI development and adoption, recruiting researchers and acquiring other AI adopters developing key AI components. This has the added benefit of restricting competitor access to valuable data / software / analytical skills, and of increasing economies of scale and scope in AI.[25] The resulting concentration in AI activity makes it easier to control deployment and restrict unsafe and unethical AI applications, but could also lead to a decline/co-opting of public AI knowledge bases and skills potentially raising barriers to AI deployment in other sectors and organizations, and less competitive and more fragile AI markets with higher concentration of activity in small number of organizations that become an attractive target for malicious actors.

4 Social complexity

Who is to blame for unsafe and disruptive AI deployment modes? In this section, I suggest that the answer could be 'everyone'. The reason for this is that individuals make, as consumers and

[25] In other words, it limits the scope for imitation of AI experiments discussed in Section 2.

tax-payers, decisions that influence AI deployment in other sectors and public service domains.[26] If they are oblivious or indifferent to the costs of those decisions for other individuals, this might create situations of extreme AI deployment. Those with superior power and influence could even seek to manipulate deployment and make others bear the brunt of AI risks, increasing inequality.

4.1 Desirable scenario

In the desirable scenario, consumers choose products and services with a level of AI intensity in production that takes into account not only price and convenience, but also the impacts on employment and working condition in the sectors where AI is being deployed, as well as other hidden costs such as declines in safety, increases in market power in the supply side or labour market demand side etc. They opt for those products and services that generate improvements in productivity sufficient to compensate those who lose-out from AI disruption.

4.2 Deviations

If the products and services supplied by industries that follow extreme models for AI deployment are cheaper or more convenient, consumers might opt for them even if they create disruption in the sectors where they are adopted, or they result in hidden costs due to declines in safety, increases in market power, rights infringements and externalities such as increased social spending in response to lower salaries and rising unemployment. These behaviours are more likely if individuals are selfish, if they are unaware of the working conditions in sectors with extreme AI deployment, or if they are myopic or uninformed about the indirect impacts of extreme AI deployment. The fact that their own working conditions are affected by consumers in other sectors making decisions in a similar way turns this situation into a social dilemma (the scenario is described more formally in the Mathematical Annex, and illustrated in Figure 5). A key idea here is that if these individuals had coordinated their consumption choices, they might have opted for less extreme, more balanced models for AI deployment that avoided or compensated for some of the disruption brought about by AI systems.

Individuals might seek to create barriers to extreme adoption of AI in their own sector. For example, their sector could be highly concentrated and therefore less responsive to consumer demand for AI products and services, its workers might be able to impose regulatory barriers to the adoption of AI , or they might have the knowledge and influence to steer AI research towards AI applications in other sectors but not theirs.

If (as might be expected) individuals' influence correlates with their political and economic power, then those in more powerful and wealthier social groups could manipulate AI deployment for their benefit at the expense of weaker, disadvantaged and disenfranchised groups. In line with this, some sectors experiencing faster AI deployment, such as 'gig economy' platforms and e-commerce, tend to employ workers who are less educated, affluent and politically influen-

[26] Their demand determines the level of AI deployment ('intensity') in the sectors they consume from.

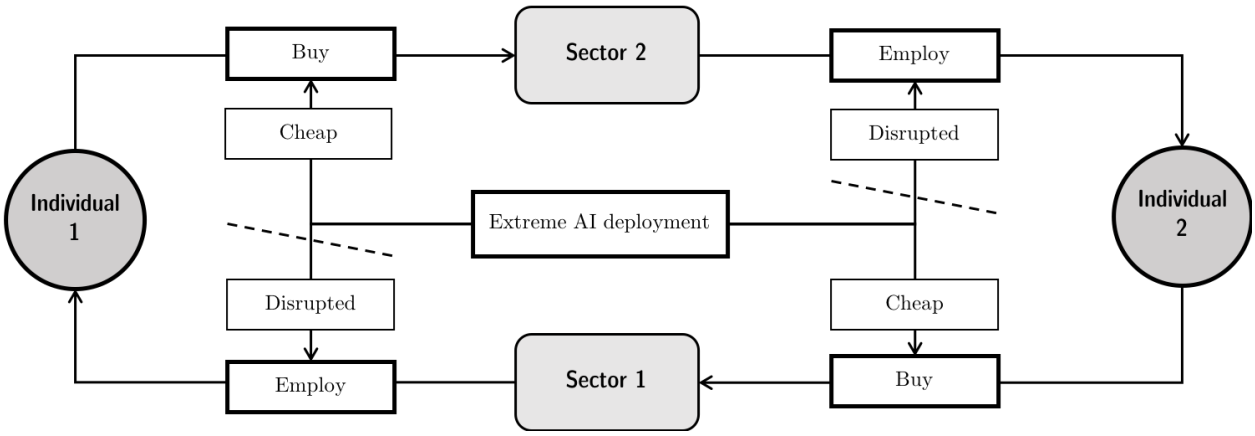

Figure 5: This figure illustrates the social dilemma between individuals who buy cheap AI-based products from other sectors and in doing so degrade the working conditions of other individuals employed there. The dashed lines indicate deployment trajectories that powerful individuals will try to prevent to avoid disruption in their own sector. The text in the boxes could be modified to represent a situation where AI is being deployed in public service domains.

tial, while deployment of AI in sectors that employ more educated and wealthy groups such as professional services or education are experiencing slower, more balanced deployment.[27]

Individuals who derive most or all their income from capital and are not altruistic or aware and responsive to indirect/hidden AI costs have incentives to encourage extreme AI deployment *across the economy*, since this gives them high consumption benefits with no workplace costs in terms of unemployment or worsened work conditions. Further, we might expect extreme AI deployment to be more profitable for them than balanced deployment, at least in the short term. Given strong concentration in capital income, this group has the economic and political influence to steer AI deployment in an extreme direction.

4.3 Public sector deployment

The discussion above could be extended to social choices about the deployment of AI in the delivery of public services. In this case, tax payers select the levels of AI deployment in the public services that other groups use. Extreme deployment creates costs for users in terms of AI errors, less ability to challenge algorithmic decisions etc., and benefits for tax-payers in terms of lower taxes. As before, uncoordinated decision-making could lead to extreme levels of adoption, and if powerful social groups steer AI deployment towards extreme models in the public services most used by weaker groups, increasing inequality.[28] This conclusion echoes concerns about the extreme and careless deployment of AI in public services involving vulnerable groups such as mi-

[27] It is of course difficult to disentangle the political economy forces I am focusing on here from other factors such as differences in the skill composition, potential AI impacts and complementary factors that also affect the deployment of AI in different sectors. This seems a fertile area for future research.

[28] A variant of this model is where majorities opt for the deployment of AI systems that create costs for minorities, for example if they have been trained on biased data reflecting a history of prejudice or inequality.

norities (policing, criminal justice system and immigration) and poorer and less educated groups (e.g. social care).

5 Temporal complexity

What decisions about AI systems being made today could take us down deployment trajectories that we we might regret in the future? The discussion so far has given some cause for concern. It suggests that uncertainty in AI complements, information asymmetry thickets, social dilemmas and uneven distribution of power between social groups could result in AI deployments that are mediocre, unsafe, abusive and unjust.

An important question is how easy - or hard - will it be to shift the direction of deployment in response to new information about its benefits and costs, and their distribution. Today's experiments would feel less risky if there was scope to overturn them, were they found to be a failure.

5.1 Desirable scenario

In a desirable scenario for AI deployment, decisions about what research, technology and business trajectories to pursue today take into account future benefits and costs for all agents, including dimensions of performance that might seem less relevant today but could prove important in the future, and changes in social behaviour and cultural attitudes that unfold as powerful AI systems are deployed in the economy and society.

Multiple social groups, actors, and disciplinary and national perspectives are incorporated into the formulation and execution of AI R&D agendas. AI deployment involves an active monitoring of irreversible commitments, sunk costs and points of no return in order to avoid scenarios that reduce diversity, competition and the scope for future choices. There is an active effort to maintain pluralism in the portfolio of AI R&D activities and business models that are explored, acknowledging uncertainty about what paths will prove rewarding and which ones will not. The loss of efficiency due to reduced scales of activity and the need to address fragmented markets and varied user needs is accepted as the cost of keeping societal AI options open, and learning from their parallel exploration.

5.2 Deviations

Uncertainty about the benefits and costs of different AI trajectories, sunk costs in deployment (specially for more deliberative or patient models of deployment that deliver benefits over longer time horizons), path-dependence in scientific and technological knowledge (the fact that the costs of switching technological trajectories accumulate over time), and network effects in AI industries and platforms create the risk of lock-in to AI trajectories which could eventually be found inferior

but hard to deviate from).[29] I go through these sources of inertia in turn (the Mathematical Annex formalizes these ideas, and 6 summarizes them).

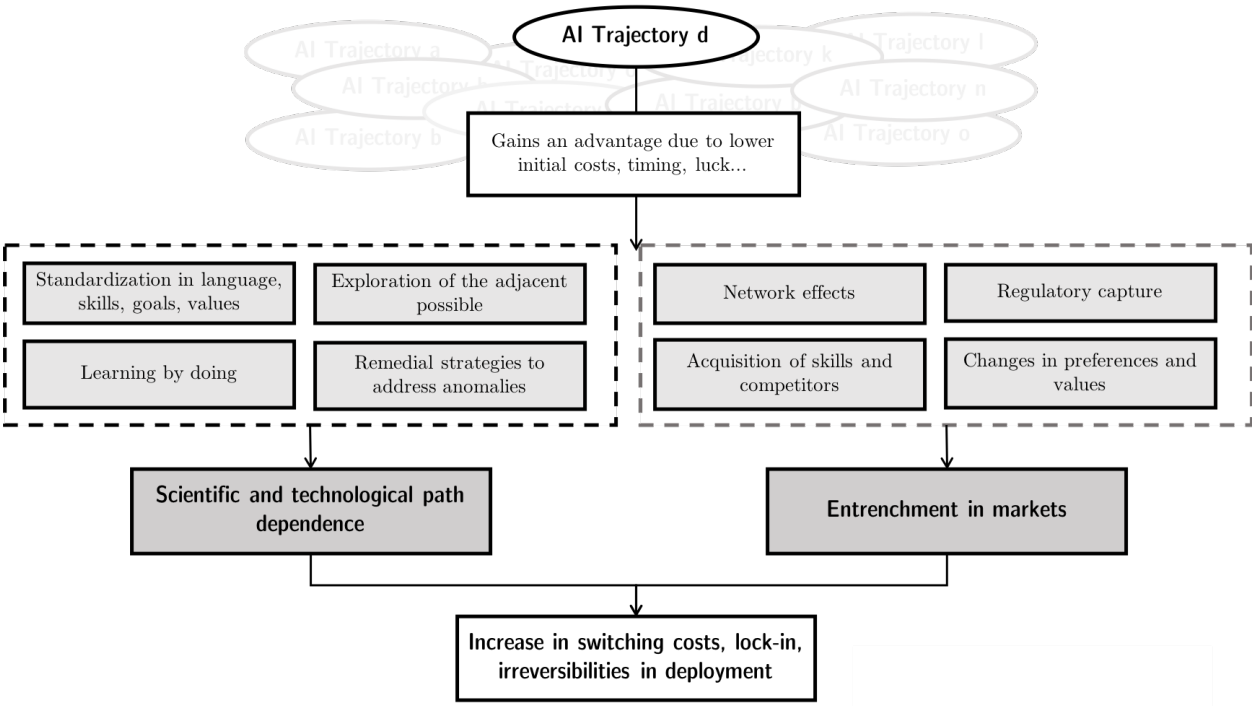

Figure 6: Sources and outcomes of path dependence in AI trajectories.

Path-dependence in science and technology

Scientific and technological progress takes place inside paradigms [61, 22] that are populated and animated by a communities of scientists/ technologists with a shared language, tools, methods and worldview (including agreement about what problems and questions are more important or interesting). This standardization improves communication and facilitates the accumulation of knowledge, including through learning-by-doing where the effectiveness of methods and technologies improves over time as their glitches are ironed out, and their complementary combinations of inputs are discovered [62]. This also means that new technological opportunities are identified through an exploration of the adjacent possible: a new advance opens up additional opportunities that build on it [63]. Pursuing these opportunities through incremental innovations along a technological trajectory is less risky than trying to open up completely new trajectories, and may indeed be the only option if new components of a technological architecture have to be integrated with existing ones. All these factors explain why alternative scientific and technological approaches tend to suffer a 'liability of the new' when they face the incumbents.

[29] Following [22], I define technological trajectories as the manifestation of an AI technological paradigm a particular combination of technologies and AI components to achieve valued technological and organizational objectives. I discuss further the components of the current AI paradigm in Sub-section 5.2

The current state of AI R&D development echoes some of these dynamics: 'normal AI science' involves the development of powerful deep learning and reinforcement learning algorithms that compete against each other and humans in benchmark datasets [28]. Although this paradigm has achieved much, it also produces AI systems that are brittle, opaque, reliant on large volumes of labelled data and computation and perhaps more focused on human displacement than augmentation [43]. These limitations are being addressed through investments in AI complements such as controlled environments for the deployment of AI that restrict the amount of new situations it is likely to face, additional data collection and labelling through crowdsourcing and Generative Adversarial Networks and investments in specialized hardware and more powerful cloud computing infrastructures.[30] An abundant supply of these complementary inputs and infrastructures, together with the skills of thousands of ML and AI researchers trained in the worldview, methods and preferred metrics of the dominant deep learning paradigm make it hard to challenge by alternative approaches that could address some of its limitations.[31]

Entrenchment in markets

Previous sections have described multiple forces pushing towards concentration in AI markets, including hard to imitate, subtle AI deployments, the desire to internalize externalities from experimentation and reduce transaction costs in AI markets with strong information asymmetries, control over important AI components, and rent-seeking by influential actors. This concentration could become entrenched for several reasons.

First there are network effects where an increase in the number of users or suppliers of complementary goods in a platform such as an app store makes it more attractive [64]. As with technology standards, challenging these platforms requires persuading a critical mass of users to coordinate their migration to a new venue. AI systems that increase a platform's ability to extract information from its user base, make better recommendations, filter information more effectively, or improve the efficiency of its processes could make the position of established internet platforms even harder to contest than it is already [10]. Profitable AI-driven organizations also have the resources to attract research talent, acquire potential competitors and lobby government for beneficial policies and regulations, further cementing their dominance.

Changes in individual preferences, attitudes and habits also increase the costs of switching away from the dominant players and business models by altering the dimensions of quality along which products and services are evaluated: for example, if consumers become less concerned about privacy or unsettled by highly accurate personalized recommendations, this can decrease the attractiveness of AI trajectories that are less reliant on the analysis of personal data. If consumers grow accustomed to algorithmic errors and glitches, or to interact with inscrutable black box algo-

[30] One could even argue that demand for public sector innovations such as the Universal Basic Income to tackle the risk of mass unemployment is a 'policy complement' to labour displacing AI systems developed in this paradigm.

[31] Having said this, a growing recognition of the brittleness and inefficiency in Deep Learning, including anomalies such as 'adversarial examples' is renewing interest in AI approaches involving causal concepts, common sense and logic.

rithms, this will reduce demand for improvements in algorithmic safety and interpretability.[32]

5.3 Outcomes

The discussion above implies that AI deployment has first-mover advantages: once a scientific and technological trajectory becomes established, and AI leaders gain dominance, it might be difficult to challenge them.

An awareness of this dynamic motivates strong investments on AI development by corporations and nations seeking to gain a position of dominance in AI markets, and to shape AI's technological trajectory in a way that is aligned with their own capabilities and values. This thinking underpins the unfolding 'AI global race' between territories such as the US, China or the EU whose visions for AI diverge in terms of the role of the private and public sector, the importance of privacy and explainability, and willingness to apply AI systems in sensitive domains such as military applications and for government surveillance [65].

This race could become one to the bottom in terms of safety, user rights, labour augmentation and impacts in productivity if flawed systems and practices become locked-in, also resulting in AI systems of limited or risky applicability in other sectors where dimensions of performance such as explainability or safety are more important. In other words, an undue narrowing of the trajectory of AI deployment could limit its generality and ultimately, its benefits for humanity.

6 Conclusion: Towards a New Science (and Policy) of the Artificial

In this essay I have argued that even though the direct economic impacts of AI (cheaper predictions and more decisions) are simple, its impacts are anything but. Complementarities in inputs, information asymmetries in markets, social dilemmas in deployment and path dependence over time could limit the benefits of AI to a small number of organizations, sectors and social groups, reduce safety, abuse rights, concentrate power and increase inequality in ways that could be socially and politically unsustainable [13, 53, 12].

Avoiding these negative scenarios requires a new agenda of research and policy to understand and manage the economic complexities that powerful AI systems bring. Following Herbert Simon's terminology, I refer to them as *New Sciences (and Policies) of the Artificial* [2]. I conclude by reviewing some principles that should inform this agenda.

First, it is important to recognize that AI is not a neutral technology and that some trajectories of deployment will be societally preferable over others based on the scale of their impact ansd its distribution over sectors, organizations, social groups and time. It is necessary to *identify and strengthen those societally beneficial trajectories of AI R&D*. Drawing on some of the issues highlighted above, this includes supporting streams of research to improve theoretical understandings of AI

[32] Ultimately, highly pro-active, personalized recommendations from AI systems could even decrease the exercise of individual agency on which the functioning of markets and democratic institutions is predicated. This raises important questions about cultural evolution and its drivers and inter-temporal aspects beyond the scope of this essay.

systems and to develop AI systems that are safer, more explainable and more conducive to labour augmentation, and supporting innovation missions that encourage the deployment of AI in new sectors where it could create significant public and social benefits, kick-starting processes of experimentation and learning that speed up deployment. This can also help countervail commercial biases towards the development of AI systems that are opaque, narrow and labour displacing. Delivering this agenda will require suitable incentives for researchers and AI adopters (for example, programmes for AI diffusion in business that encourage labour augmenting AI technologies), improving the governance of AI research and a better evidence base about the pace and direction of AI deployment based on detailed, timely data itself analyzed using new sources of data and AI and ML methods [66, 67].

Second, there is the need for *experimentation and evidence* to reduce current levels of uncertainty about the (broadly defined) economic impacts of AI and its complements. This involves a systematic programme of experimentation in a variety of organizational contexts, using rigorous methods and paying special attention to indirect impacts and side-effects of AI deployment [68]. This sort of analysis will be particularly important in high-stakes domains where there are justifiably large barriers to experimentation, and in collective intelligence environments involving complex collaborations between many organizations [8]. Comparing current organizational performance with what could be achieved through the judicious deployment of AI systems can take us beyond over-pessimistic views of the impacts of AI that risk entrenching a status-quo dominated by human decision-makers that are not without their failures and biases [30]. The results of these studies should be widely and consistently shared to maximize collective learning about the opportunities and risks of AI systems in different sectors, and to encourage diffusion of good practices and coordinate the supply of complementary inputs, in particular skills.

Third, it is critical to ensure *transparency and compliance* in the adoption of AI systems to reduce information asymmetries in AI markets. This involves regulatory systems that clearly identify which AI applications are permitted and which are not, monitoring systems to ensure compliance with those rules, and systems to increase public awareness of how AI systems are deployed in different organizations and platforms to help consumers make better-informed decisions (this information should be accompanied by complementary policies to enable consumer exit and voice, such as data portability and rights to explanation [6]). A thorny issue here will be that of balancing safety and fairness with the freedom to experiment so as to avoid stifling innovation, particularly in sectors and domains where the incentives for this are weaker in the first place. A stronger evidence base about the nature and extent of AI impacts and their complementary inputs following the experimentation and evidence principle could help delineate more effectively what are the areas where AI innovation can be permisionless, and those where it should be restricted [69]. All these efforts should build on and complement ongoing efforts to embed ethical principles into AI that are necessary but not sufficient, without verification and penalties, to ensure well-functioning AI markets [70].

Fourth, and in line with the notion of Rawlsian AI deployment that informs this essay, social

dilemmas in AI deployment can only be avoided by ensuring *coordination and social solidarity in AI deployment*: building on the points above, this requires supporting directions of AI R&D agreed to be desirable, encouraging experiments to verify their impacts and minimize their negative side effects, and designing and enforcing regulations about AI deployment. The spatial, sectoral and social distribution of AI impacts should be closely monitored to inform compensatory interventions aimed at ensuring that the benefits and costs of AI deployment are fairly shared [13]. All this needs to be informed by a programme of broad-based public engagement that reflects the generality of AI and the far-reaching nature of its impacts.

Fifth and last, *diversity in AI trajectories* should be preserved to avoid a premature lock-in to inferior paths of deployment. This will require active monitoring of the scientific, technological and economic landscape to identify sources of concentration, homogeneity and fragility, and proactively intervening to preserve diversity through targeted streams of R&D funding, competition policy and regulation. Any irreversibilities and sources of lock-in that are identified should be openly and publicly debated and, as much as possible, empirically explored through experiments and test-beds to minimize the risk of regretting the trajectories of development that are eventually pursued.

Following these principles will require significant institutional innovations involving new actors or networks of actors to coordinate research and policy activities, and new systems to collect and analyze data and design and implement interventions, including via AI systems [71]. These policy directions should be shaped by social agreements about what is useful, desirable and fair that can only reached through sustained public discussion and social learning where economic analyses, experiments and data will be an important input among many. One particular challenge here will be to find ways to align the motivations and interests of countries with different, potentially conflicting visions for AI deployment. One could reformulate Rawls' original position around countries instead of individuals [24], and consider what principles and institutions should govern the global deployment of AI in a way that is just, and accepted by all [72]. Such analysis goes beyond the scope of this essay, but it is an urgent one to undertake.

Whatever shape it takes, the process of collective self-discovery that lays ahead [73] could work as a mirror version of the Turing test where different human societies learn about themselves - their values, goals and capabilities - through their responses to the opportunities and challenges opened up by the powerful AI systems they have created.

References

[1] W. Ross Ashby. Requisite variety and its implications for the control of complex systems. In *Facets of systems science*, pages 405–417. Springer, 1991.

[2] Herbert A. Simon. *The sciences of the artificial*. MIT press, 1996.

[3] Richard R Nelson and G Sidney. Winter. 1982. an evolutionary theory of economic change, 2005.

[4] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction Machines: The simple economics of*

artificial intelligence. Harvard Business Press, 2018.
[5] Matt Taddy. The Technological Elements of Artificial Intelligence. Technical report, National Bureau of Economic Research, 2018.
[6] Eric A Posner and E Glen Weyl. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press, 2018.
[7] Iain M. Cockburn, Rebecca Henderson, and Scott Stern. The Impact of Artificial Intelligence on Innovation. Technical report, National Bureau of Economic Research, 2018.
[8] Geoff Mulgan. *Big Mind: how collective intelligence can change our world*. Princeton University Press, 2017.
[9] Ajay K Agrawal, Joshua Gans, and Avi Goldfarb. Introduction to" economics of artificial intelligence". In *Economics of Artificial Intelligence*. University of Chicago Press, 2017.
[10] Jason Furman and Robert Seamans. AI and the Economy. SSRN Scholarly Paper ID 3186591, Social Science Research Network, Rochester, NY, May 2018.
[11] Timothy F. Bresnahan and Manuel Trajtenberg. General purpose technologies 'Engines of growth'? *Journal of econometrics*, 65(1):83–108, 1995.
[12] Manuel Trajtenberg. AI as the next GPT: a Political-Economy Perspective. Technical report, National Bureau of Economic Research, 2018.
[13] Joseph E. Stiglitz and Anton Korinek. Artificial Intelligence, Worker-Replacing Technological Change, and Income Distribution. *NBER Chapters*, 2017.
[14] Philippe Aghion, Benjamin F. Jones, and Charles I. Jones. Artificial Intelligence and Economic Growth. Technical report, National Bureau of Economic Research, 2017.
[15] Daron Acemoglu and Pascual Restrepo. Artificial Intelligence, Automation and Work. Technical report, National Bureau of Economic Research, 2018.
[16] Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete Contracting and AI Alignment. *SSRN Electronic Journal*, 2018.
[17] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, and Bobby Filar. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.
[18] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.
[19] DeepMind Safety Research. Building safe artificial intelligence: specification, robustness, and assurance, September 2018.
[20] Paul A. David. Clio and the Economics of QWERTY. *The American economic review*, 75(2):332–337, 1985.
[21] Paul A David. The dynamo and the computer: an historical perspective on the modern productivity paradox. *The American Economic Review*, 80(2):355–361, 1990.
[22] Giovanni Dosi. Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Research policy*, 11(3):147–162, 1982.
[23] Philippe Aghion, Paul A. David, and Dominique Foray. Science, technology and innovation for economic growth: linking policy research and practice in 'STIG Systems'. *Research policy*, 38(4):681–693, 2009.
[24] John Rawls. *A theory of justice*. Harvard university press, 2009.
[25] John Markoff. *Machines of loving grace: The quest for common ground between humans and robots*. HarperCollins Publishers, 2016.
[26] John Seely Brown and Paul Duguid. *The Social Life of Information: Updated, with a New Preface*. Harvard Business Review Press, 2017.
[27] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. A

brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*, 2017.
[28] Miles Brundage. Modeling Progress in AI. In *AAAI Workshop: AI, Ethics, and Society*, 2016.
[29] AI Index. The Artificial Intelligence Index: 2017 Annual Report. Technical report, 2017.
[30] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2017.
[31] Colin F Camerer. Artificial intelligence and behavioral economics. page 33.
[32] Daniel Kahneman and Patrick Egan. *Thinking, fast and slow*, volume 1. Farrar, Straus and Giroux New York, 2011.
[33] Andrew McAfee and Erik Brynjolfsson. *Machine, platform, crowd: Harnessing our digital future*. WW Norton & Company, 2017.
[34] Joel Klinger, Juan C. Mateos-Garcia, and Konstantinos Stathoulopoulos. Deep Learning, Deep Change? Mapping the Development of the Artificial Intelligence General Purpose Technology. SSRN Scholarly Paper ID 3233463, Social Science Research Network, Rochester, NY, August 2018.
[35] Ajay Agrawal, John McHale, and Alex Oettl. Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth. Working Paper 24541, National Bureau of Economic Research, April 2018.
[36] Nicholas Bloom, Charles I. Jones, John Van Reenen, and Michael Webb. Are ideas getting harder to find? Technical report, National Bureau of Economic Research, 2017.
[37] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.
[38] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
[39] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.
[40] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.
[41] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.
[42] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
[43] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
[44] Joel Mokyr et al. *The gifts of Athena: Historical origins of the knowledge economy*. Princeton University Press, 2002.
[45] Jonathan Haskel and Stian Westlake. *Capitalism without capital: the rise of the intangible economy*. Princeton University Press, 2017.
[46] Juan Mateos-Garcia. To Err is Algorithm: Algorithmic fallibility and economic organisation. 2017.
[47] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. The simple economics of machine intelligence. *Harvard Business Review*, 17, 2016.
[48] Nicholas Bloom, Luis Garicano, Raffaella Sadun, and John Van Reenen. The distinct effects of information technology and communication technology on firm organization. *Management Science*, 60(12):2859–2885, 2014.
[49] Hasan Bakhshi, Albert Bravo-Biosca, and Juan Mateos-Garcia. The analytical firm: Estimating the effect of data and online analytics on firm performance. Technical report, Nesta Working Paper 14/05, 2014.

[50] Melanie Arntz, Terry Gregory, and Ulrich Zierahn. The risk of automation for jobs in OECD countries. 2016.
[51] David J Deming. The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, 132(4):1593–1640, 2017.
[52] Carl Benedikt Frey and Michael A. Osborne. The future of employment: how susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017.
[53] David Autor and Anna Salomons. Is automation labor-displacing? Productivity growth, employment, and the labor share. *Brookings Papers on Economic Activity*, 2018.
[54] James E. Bessen. How computer automation affects occupations: Technology, jobs, and skills. 2016.
[55] David J Deming and Kadeem L Noray. Stem careers and technological change. Technical report, National Bureau of Economic Research, 2018.
[56] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *arXiv preprint arXiv:1807.05307*, 2018.
[57] Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. The Flash Crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.
[58] Jack Stilgoe, Richard Owen, and Phil Macnaghten. Developing a framework for responsible innovation. *Research Policy*, 42(9):1568–1580, 2013.
[59] Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*, 2018.
[60] Nick Bostrom. Strategic implications of openness in ai development. *Global Policy*, 8(2):135–148, 2017.
[61] Thomas S. Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.
[62] Kenneth J Arrow. The economic implications of learning by doing. *The review of economic studies*, 29(3):155–173, 1962.
[63] W. Brian Arthur. *The nature of technology: What it is and how it evolves*. Simon and Schuster, 2009.
[64] Carl Shapiro, Shapiro Carl, and Hal R. Varian. *Information rules: a strategic guide to the network economy*. Harvard Business Press, 1998.
[65] Avi Goldfarb and Daniel Trefler. AI and International Trade. Technical report, National Bureau of Economic Research, 2018.
[66] Robert Seamans and Manav Raj. Ai, labor, productivity and the need for firm-level data. Technical report, National Bureau of Economic Research, 2018.
[67] J. Klinger, J. Mateos-Garcia, and K. Stathoulopoulos. Deep learning, deep change? Mapping the development of the Artificial Intelligence General Purpose Technology. *arXiv preprint arXiv:1808.06355*, 2018.
[68] Kate Crawford and Ryan Calo. There is a blind spot in ai research. *Nature News*, 538(7625):311, 2016.
[69] Adam Thierer. *Permissionless innovation: The continuing case for comprehensive technological freedom*. Mercatus Center at George Mason University, 2016.
[70] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. Does acm's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 729–733. ACM, 2018.
[71] Ryan Calo. Artificial intelligence policy: A primer and roadmap. *UCDL Rev.*, 51:399, 2017.
[72] Stephen Cave and Seán SOhÉigeartaigh. An ai race for strategic advantage: Rhetoric and risks. In *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*, 2018.
[73] Ricardo Hausmann and Dani Rodrik. Economic development as self-discovery. *Journal of*

development Economics, 72(2):603–633, 2003.
[74] Karl Popper. *The poverty of historicism*. Routledge, 2013.
[75] Jeffrey D Sachs, Seth G Benzell, and Guillermo LaGarda. Robots: Curse or blessing? a basic framework. Working Paper 21091, National Bureau of Economic Research, April 2015.

Mathematical annex

The sub-sections here explore mathematically the issues raised by different types of complexity considered over the essay.

Organizational Complexity

Deploying AI in a firm

Consider a firm f using j components p to produce a good it sells at a unit price. This includes specifics skills and forms of capital, processes, organizational forms and business models. For example, p_1 could refer to the number of engineers, p_2 to the level of employee empowerment, p_3 to the investment in cloud computing etc.

Each of the components that the firm uses have a risk $r_i \epsilon (0,1)$ associated to its novelty. Riskier components bring potentially higher benefits B because they help the firm stand out in the market. However, they also generate failures F for example when they cause accidents and errors, create legal liabilities and reputational damages etc.[33] This means that $r_i > r_j \implies B_i > B_j$. A firm will only consider using a novel, risky component if there is an upside. Components also have a cost c_i.

Based on this, the profits of a firm π_f using component mix P will be:

$$\pi_f = \Pi_{i=0}^{j}(1-r_i)B_i - \Pi_{i=0}^{j} r_i F_i - \sum c_i \tag{1}$$

Now let us assume that a new AI system becomes available. To realize its benefits, the firm needs to acquire a complementary component (AI complement) p_{AI} with risk r_{AI}. The firm will adopt the new AI component if its expected impact in profits is positive, that is, if:

$$\frac{\Delta B}{\Delta F}\frac{(1-r_{AI})}{r_{AI}} > \frac{\Pi_{i=0}^{j}(r_i)}{\Pi_{i=0}^{j}(1-r_i)} + c_{AI} \tag{2}$$

The ratio of expected benefits to risks (from failure) from integrating an AI component need to offset the ratio of benefits to risks of the initial situation plus the costs of implementation. It is more likely that the component will be acquired if its benefits are higher, if its risk is lower, if it is incremental (the difference between its risks and the risks of the status is lower), and its costs are lower. Sectors where these conditions hold will have more incentives to deploy AI than those where they do not.

A firm might need to acquire multiple new components $P_{AI} = \{p_{AI,1}, ..., p_{AI,k}\}$ to benefit from AI. Building on equation 2, and assuming that all of the new inputs have the same risk r_{AI}, we get:

$$\frac{\Delta B_k}{\Delta F_k}\left(\frac{1-r_{AI}}{r_{AI}}\right)^k > \frac{\Pi_{i=0}^{j}(r_i)}{\Pi_{i=0}^{j}(1-r_i)} + kc_{AI} \tag{3}$$

[33] One way to think of r_i is a the probability of successful usage.

As k increases, we expect the ratio of expected benefits to costs to decline, in part because the costs of managing implementation complexity is likely to increase with the number of new components being integrated $\frac{\delta(\frac{\Delta B)}{\Delta F}}{\delta k} > 0$.[34]. All this suggests that there are limits to the levels of ambition in AI deployment inside a firm.

Disruption in implementation

Larger and more conservative organizations where (1-r_i) (the reliability of existing components) and B_P (the benefits of the status quo) are higher and where the potential errors from existing practices (F_B) are lower have, other things being equal, less incentives to implement many AI components.

This situation is reinforced if the successful deployment of AI systems requires a substantive reorganization in current structures and processes, with (at least in the short term) a negative impact on performance. In our model, this means that $\frac{\delta r_i}{\delta k} < 0$. The costs of disruption is higher for larger and more complex organizations where j is higher.[35]

Modularity and patience in deployment

One potential strategy to manage the risks of implementation is through modularity and experimentation: experimenting with a smaller set of components $v < k$ and learning from the outcomes (here we assume that there is learning by doing so that $\frac{\delta(1-r_{AI,i})}{\delta t} < 0$ after implementation). This strategy will be more attractive if the expected benefits of learning over a period offset the discounted benefits of maximal adoption over that period.[36]

External vendors and their relevance

Some bundles of components ('modules') $P_{AI,V} = \{p_{AI,V_1}, ..., p_{AI,V_m}\}$ will be available from external vendors who specialize in their development and integration so that:

$$B(P_{AI,V})\Pi_{i=0}^{m} r_{AI,V_i} + c_{AI,V} - \rho(AI, V, f) > B(P_{AI})\Pi_{i=0}^{m} r_{AI,i} + c_{AI} \quad (4)$$

This means that the expected benefits from sourcing this module externally minus a 'relevance penalty' ρ are higher than the expected benefits of internal sourcing. The relevance penalty captures the fact that modules designed for a mass market of AI adopters will be less relevant for firms with unique needs and firms that are lead users in sectors where AI has been more recently adopted. Communicating those unique needs to the external vendor will be costly.[37]

[34]The derivative of the left side of 3 on k needs to be positive. With $R = \frac{\Delta B_k}{\Delta F_k}$, this means that $\frac{\delta R}{\delta k} + ln(1 - r_{AI}) - ln(r_{AI}) > 0$

[35]We could represent these interactions in a matrix $M_{i,j}$ where the rows i represent the new components and the columns j represent the existing ones. Each cell $x_{i,j}$ represents the interaction between the new component and existing practices (e.g. it could be $x_{i,j}\epsilon\{0, 1\}$ a penalty on the risk of r_j of p_j if x_i is adopted . The matrix will be wider for larger, older organizations than smaller, younger ones.

[36]The benefits of this strategy mirror those described by Simon in [2] and the notion of piecemeal social engineering versus utopian social engineering in [74]. Note that here I ignore the possibility that $\frac{\delta B}{\delta(nk)} > \sum_{k=0}^{n} \frac{\delta B}{\delta K}$, that is, the existence of complementarities in implementation, which would increase the incentives for maximal AI deployment.

[37]Later on, I consider principal agent situations in the communication - that is, the fact that it is costly to assess the expected benefits of external sourcing of AI modules.

Learning, imitation, hiding and competing

Experimentation with new components and modules reduces their risks for an organization by helping it to identify good practices. This learning has also ha public, collective dimension l: Learning increases over time, and with the number of experimenters E ($l = \frac{\delta(1-r_{AI,i})}{\delta t} > 0$ and $\frac{\delta l}{\delta E} > 0$).

Externalities from experimentation (the fact that a firm's experiments benefit its competitors by lowering their deployment risks $r_{AI,i}$) can however lead to under-investment in experimentation if many firms adopt a 'wait and see' strategy [11]. The extent of these externalities depends on the distance ρ between the leader and the imitator, the level of secrecy in its experimentation (which can be accomplished through *subtle* deployment), and implementation costs for the imitator if there are interdependencies between the new components and existing practices. Further, lead experimenters can identify critical components for AI deployment (such as data, infrastructure, specific skillsets etc) and secure their supply, which they can then use to hinder imitators, or charge them to enter the market.

Of course, all of the above implies that imitating an experimenter is a voluntary decision. If ΔB_{AI} is sufficiently large, deploying AI could be the only option for firms operating in (and not willing to exit) competitive markets, regardless of the risks.

Coda: Uncertainty in estimates of benefits, costs and risks

Mant of the parameters informing firms' decisions above, such as $R(AI), \Delta B, \Delta E$ or ρ will be biased estimates of real values with a confidence interval whose breadth depends on prevailing levels of technological and market uncertainty. This creates the risk of over and under-investment on AI, and of the adoption of mediocre or unsafe AI systems with disappointing upsides ΔB and unexpected downsides ΔE.

Market complexity

AI principals and agents

Consider a transaction between two actors in an AI market, P and A where one of them P is the principal, the A is the agent. In this situation, P is delegating an activity on A and A selects a suitable strategy to undertake this activity. A can choose from a set of strategies $S = \{S_1, ..., S_n\}$.[38] Each strategy S_i is associated to a tuple with outcomes for each actor k $(B_{i,k}, C_{i,k})$. The costs of a strategy include the costs of its inputs as well as other important costs in terms of accidents, algorithmic failures etc.

Subtlety in AI deployment creates an information asymmetry between principal and agent, giving the agent some discretion in the selection of S. Further, we assume that there is a misalignment in incentives between the principal and the agent. If $\pi_k = B_{i,k} - C_{i,k}$ then $\frac{\delta \pi_A}{\delta \pi_P} \leq 0$. Maximizing the benefits for the principal might not be the optimal decision for the agent, perhaps because it requires relinquishing some benefits (e.g. not exploiting all the data that has been obtained from the user in a social networking site) or incurring in more costs (implementing more stringent and costly supervision systems to reduce algorithmic errors). The specifics depend on the transaction.

When selecting her strategy, the agent also needs to take into account the possibility that the principal will detect her behaviour with probability d, resulting in a fine F.[39] I assume that d

[38] We could think of each of these as a set of inputs P_{AI} presented above, with a vector of risk R_i.

[39] This fine could be a legal fine or a market penalty, like the decision to stop transacting with A.

depends on the benefits of the transaction for P so that $\frac{\delta d}{\delta(1-\pi_P)} > 0$ (as the benefits of the transaction for P declines, she is more likely to become more suspicious of A's behaviour. F depends on the regulatory and competitive environment.

Given all this, the agent has an optimal strategy o such that:

$$o = \arg\max_{o \epsilon S}(\pi_A(o) - Fd((1 - \pi_P(o))) \tag{5}$$

Situations of higher market uncertainty/lower transparency where d is low, as well as less stringent regulatory or competitive regimes with lower F will lead to more aggressive profit maximization by the agent at the expense of the principal (in terms of lower benefits from a transaction or higher costs).

Hidden costs

AI costs might be hidden from the principal. Perhaps they are not paid by the principal but by other actors or groups in society she is meant to represent, or perhaps they only become manifest when they overcome a threshold T. This will reduce the incentives for the agent to take into account those costs when she selects S.[40]

Honest mistakes and error chains

An agent's estimates of π_A and π_P could be biased, for example because she is adopting risky and poorly understood AI components, or because she procured them from another agent A_2 that is maximizing its own π_{A2} at her expense. This increases Fd, not because the agent is behaving opportunistically, but because of the technological and organizational complexity of the AI system she is deploying, and opportunistic behaviours elsewhere in the AI value chain. The perception of these risks could lead the agent to behave more carefully (adopt a S associated to higher π_p), use less risky AI components or monitor more stringently the risk of the AI components she sources from other actors in this AI market.

Social complexity

I illustrate social complexity in AI deployment with a simple variation of the prisoner's dilemma. It involves an economy with two individuals I_1 and I_2 working in sectors, S_1 and S_2. Each individual works in one sector and consumes from the other. When doing this, she can opt for goods produced with a degree of AI intensity d. This has implications for the prices she pays and for the employment outcomes and working conditions of the individual working in the other sector. Each individual receives a work payoff $w_{i,d}$ (salary weighted by working conditions) and a consumption payoff $c_{i,d}$ (price weighted by convenience and quality). The AI adoption scenarios and pay-offs in sector S_i are:

1. **No AI deployment**: The prices of S_i products and the working conditions of its workforce are $w_{i,n}$ and $c_{i,n}$.

2. **Balanced AI deployment**: AI is deployed in a labour augmenting way: it increases the productivity of the workforce in the deploying sector without damaging its working conditions. The salary w_i for workers in the sector is $w_{i,b} > w_{i,n}$ due to the increase in productivity, while

[40] In 3, a decline in ΔE will result in an adoption of more AI components k.

consumers buying from the sector receive a consumption payoff $c_{i,b} > c_{i,n}$ due to decline in prices/increases in convenience and quality.[41]

3. **Extreme AI deployment**: The deployed AI system is labour displacing and creates mass unemployment. Those individuals who hold on to their jobs suffer degraded working conditions and are paid lower salaries. Workers in the sector have a payoff $w_{i,e} < w_{i,n}$. Consumers from the sector have a consumption payoff $c_{i,e} > c_{i,b}$ due to a strong drop in prices, improvements in quality and convenience etc.[42]

I assume that these individuals do not consider other important impacts of AI deployment such as for example, declines in product safety or increases in market power. These costs might be hidden from them, or heavily discounted because they will happen in the future. I also assume that individuals are either selfish or ignorant about the impact of their decisions: they fail to take into account the employment outcomes and working conditions of those in other sectors, or they are not aware of them.

I represent pay-offs in Table 1 with some illustrative values ($w_{i,n} = 0, w_{i,b} = 5, w_{i,e} = -10$ and $c_{i,n} = 0, c_{i,b} = 5, c_{i,e} = 10$). In line what I assumed above, $w_{b,i} > w_{n,i} > w_{e,i}$ and $c_{e,i} > c_{b,i} > c_{n,i}$.

$I_1 \setminus I_2$	**No deployment**	**Balanced deployment**	**Extreme deployment**
No deployment	I_1:(0,0)	I_1:(0,5)	I_1:(0,10)
	I_2:(0,0)	I_2:(5,0)	I_2:(-10,0)
Balanced deployment	I_1:(5,0)	I_1:(5,5)	I_1:(5,10)
	I_2:(0,5)	I_2:(5,5)	I_2:(-10,5)
Extreme deployment	I_1:(-10,0)	I_1:(-10,5)	I_1:(-10,10)
	I_2:(0,10)	I_2:(5,10)	I_2:(-10,10)

Table 1: Pay-offs for AI deployment: Each individual I_i controls, through her consumption choices, the intensity d of AI deployment in the other sector S_i. The pay-off tuple represents pay-offs for individual working in each sector S_i. The first value in each tuple represents the work payoff $w_{i,d}$, and the second value represents the consumption pay-off $c_{i,d}$. Although the social optimum is achieved when both individuals select balanced adoption strategies in the other sector, the equilibrium strategies are to select extreme adoption.

If we assume that pay-offs are fungible across individuals and activities ($\pi_{d,d} = \sum_i w_{i,d} + \sum_i c_{i,d}$), we see that the societally optimal scenario involves balanced AI deployment of both sectors, with $\pi_{b,b} = 20$. However, this scenario is unstable: without coordination, both individuals have incentives to demand cheap and convenient goods based on extreme AI from the other sector. As a result, AI is extremely deployed everywhere with $\pi_{e,e} = 0$. All individuals suffer unemployment and/or degraded working conditions but enjoy access to cheaper and more convenient goods.[43] If they could coordinate their decisions, these individuals would have opted for a deployment scenario with a stronger work-life balance (literally).[44]

[41] I assume that workers in a sector are not adverse to technological change. If that was the case, $w_{i,b} < w_{i,n}$.

[42] The increase of productivity may create new economic opportunities for displaced workers but this is by no means certain, and in the short-term they will experience disruption.

[43] However, cheaper goods may not be affordable for unemployed individuals as in [75].

[44] With these values of $w_{i,d}$ and $c_{i,d}$ both individuals are indifferent between no adoption and extreme adoption. The

Temporal complexity

Consider a situation where there are two alternative technological trajectories for AI, T_1 and T_2. Each of them has an initial deployment cost d_i (this could include investments in infrastructure, costs of adopting AI in businesses and government etc.). In each period t, the trajectory also incurs a cost c_i which includes the labour costs of AI researchers, engineers and supervisors, upkeep of the data infrastructure and models and costs from algorithmic errors, disruption in labour markets etc. Each period, the trajectory also generates benefits b_i in terms of improved productivity.[45] What technology will be adopted? If we assume that future returns are perpetual and discounted at a rate r, then T_1 will be selected if:

$$\frac{(b_1 - b_2) - (c_1 - c_2)}{r} > d_2 - d_1 \tag{6}$$

That is, the net present value of T_1's returns compared to T_2 have to be superior to its deployment costs at the onset. This is the same as:

$$r > \frac{(b_2 - b_1) - (c_2 - c_1)}{d_2 - d_1} \tag{7}$$

If the discount rate is sufficiently high (for example, if there are high levels of uncertainty about the future leading to high discounts in future returns, or a perception of 'winner-takes-all' in AI markets), AI technologies that are cheaper to deploy (low d_i) might be selected even if they are inferior on a period-by-period basis.[46] For example, if equation 7 holds and T_1 is adopted, there will be no incentives to switch to the alternative T_2 for as long as $\frac{(b_2-b_1)-(c_2-c_1)}{r} > d_2$ which is bound to hold given equation 6. Changing trajectories is even harder if there is a path-dependence - a switching cost l_i that accumulates over the time that a technology is installed. This could happen if AI researchers and engineers learn skills that are specific to this AI technological trajectory, or if users become accustomed to the platforms that use this modality of AI and its business models (such as for instance trading personal data for 'free' goods and services). In that case, in period t it only makes sense to switch to T_2 if $\frac{(b_2-b_1)-(c_2-c_1)}{r} > d_1 + \sum_{i=0}^{t} l_i$.

We can even imagine scenarios where T_1 becomes progressively worse - it is shown to be a technological dead-end with hidden costs - yet there is no economic reason to switch to T_2. As long as $\frac{\delta(c_1-b_1)}{\delta t} < l_1$, the costs of switching from this deteriorating AI trajectory to T_2 *increase* over time.

result would change with different pay-offs - say, if individuals place a premium on work over consumption, or if they experience loss-aversion so that losses in $w_{i,d}$ outweigh gains in $c_{i,d}$.

[45] This set-up mirrors the organizational complexity model, with ΔB captured in b_i and ΔE and c_i in c_i.

[46] Here we could assume that human-displacing and less safe systems might enjoy an advantage over human-augmenting and safe systems because they are simpler and cheaper to deploy, as discussed in Sections 2 and 3.