Berkeley
UNIVERSITY OF CALIFORNIA
BERKELEY ROUNDTABLE ON
THE INTERNATIONAL ECONOMY

BRIE Working Paper
2020-5

Governing AI:
Understanding the Limits, Possibility, and Risks of AI in an Era of Intelligent Tools and Systems

John Zysman and Mark Nitzberg

Acknowledgments: This research was funded in part by the Ewing Marion Kauffman Foundation with additional support from CITRIS, The Trigger Project of the European Commission, and others. The contents of this publication are solely the responsibility of the authors.

BRIE Working Paper # 2020-5
August 2020
(Draft for comments and discussion)

Governing AI:

Understanding the Limits, Possibility, and Risks of AI in an Era of Intelligent Tools and Systems

(zysman.john@gmail.com)
(comments welcome)

John Zysman
Professor Emeritus
Department of Political Science
Business Roundtable on the International Economy (BRIE)
University of California, Berkeley

&

Mark Nitzberg
Executive Director
Center for Human-Compatible Artificial Intelligence (CHAI)
Berkeley AI Research (BAIR)
University of California Berkeley

Abstract: In debates about artificial intelligence (AI), imaginations often run wild. Policy-makers, opinion leaders, and the public tend to believe that AI is already an immensely powerful universal technology, limitless in its possibilities. However, while machine learning (ML), the principal computer science tool underlying today's AI breakthroughs, is indeed powerful, ML is fundamentally a form of context-dependent statistical inference and as such has its limits. Specifically, because ML relies on correlations between inputs and outputs or emergent clustering in training data, today's AI systems can only be applied in well-specified problem domains, still lacking the context-sensitivity of a typical toddler or house-pet. Consequently, instead of constructing policies to govern artificial general intelligence (AGI), decision-makers should focus on the distinctive and powerful problems posed by narrow AI, including misconceived benefits and the distribution of benefits, autonomous weapons, and bias in algorithms. AI governance, at least for now, is less about managing super-intelligent systems than about managing those who would create and deploy them and supporting the application of AI to narrow, well-defined problem domains.

Specific implications of our discussion are as follows:

- AI applications are part of a suite of intelligent tools and systems and must ultimately be regulated as a set. Digital platforms, for example, generate the pools of big data on which AI tools operate and hence, the regulation of digital platforms and big data is part of the challenge of governing AI. Many of the platform offerings are, in fact, deployments of AI tools. Hence, focusing on AI alone distorts the governance problem.

- Simply declaring objectives—be they digital privacy, transparency, or avoiding bias—is not sufficient. We must decide what the goals actually will be in operational terms.
- The issues and choices will differ by sector. The consequences, for example, of bias and error will differ from a medical domain or a criminal justice domain to one of retail sales.
- The application of AI tools in public policy decision making, in the design of transport or waste disposal or policing, or in a whole variety of domains, requires great care. There is a substantial risk of confusing efficiency with public debate about what the goals should be in the first place. Indeed, public values evolve as part of social and political conflict.
- The economic implications of AI applications are easily exaggerated. Should public investment concentrate on advancing basic research or on diffusing tools, user interfaces, and the training needed to implement them?

As difficult as it will be to decide on goals and a strategy to implement the goals of one community, let alone regional or international communities, any agreement that goes beyond simple objective statements is very unlikely.

Governing AI:

Understanding the Limits, Possibility, and Risks of AI
in an Era of Intelligent Tools and Systems

John Zysman[i] and Mark Nitzberg[ii]

August 2020

> *"….we need to reconceptualize it [AI – Deep learning]: not as a universal solvent, but simply as one tool among many, a power screwdriver in a world in which we also need hammers, wrenches, and pliers, not to mentions chisels and drills, voltmeters, logic probes, and oscilloscopes.*
>
> *In perceptual classification, where vast amounts of data are available, deep learning is a valuable tool; in other, richer cognitive domains, it is often far less satisfactory."*
>
> - Gary Markus (p. 18, 2018) [iii]
>
> *"Intelligent tools are diffusing through our economies and society. Some of the developments are powerfully changing how our economies work and how we live our lives. Some of the purported developments are simply hype. Amid the froth, many believe that our current social and economic arrangements will be swept aside and, at the extreme—that we will become the metaphorical "pets" of super-intelligences. Others, ourselves included, assert that the world is ours to create. That is easy to assert but difficult to demonstrate and harder still to implement."*
>
> - John Zysman, Martin Kenney, Laura Tyson (p. 2, 2019)

Governance in this era of intelligent tools and systems requires stepping beyond the hype about particular tools and the despair that overblown claims about them can engender. The phrase "intelligent tools and systems" points, therefore, to the toolbox, not individual tools. The toolbox itself is constantly expanding, and the tools are constantly gaining power. Importantly, given the suite of intelligent tools, governance issues about particular tools cannot usefully be dealt with entirely in isolation. Hence the debate about artificial intelligence (AI) governance must be about the *set* of intelligent tools and how they relate to each other.

Consider what is left out of the conversation if we only discuss AI. Take, for example, two-sided digital platforms (hereafter termed digital platforms). Digital platforms, key tools in the box, are usually treated separately from AI and big data. But digital platforms are very much the nexus of both. Platforms amass the big chunks of data required for the effective deployment of AI tools. In turn, AI tools, which depend on big data, give power to the creators of these platforms. Consequently, digital platforms, one may argue, are the most fundamental of the tools in the tool box, often creating the framework for developing big data and applying statistical tools such as deep learning to that data. Platforms "are an emblem and embodiment of the digital era just as factories were of the industrial revolution" (Bearson, Kenney, and Zysman, p. 4, 2019; Kenney, Bearson, and Zysman, p. 2, 2019) Thus, focusing only on AI would be like focusing only on steam engines and the rules for steam engines when factories emerged. Of course, digital

platforms, data, and AI applications all run on and depend on the ever-increasing power of digital infrastructure and computing in the cloud. So, governing platforms, a significant debate in itself, is entangled within any discussion about governing data and AI. Seen from that vantage, AI is simply a part of the challenge of governing the platform economy—the economy of intelligent tools and systems.

To clarify our position, the core story about governance in a digital era is not about the rules of particular tools, but the challenges represented by a new tool box. That toolbox includes big data, powerful computing capacity, algorithms, and software in general—not only that driving machine learning in its various forms. AI, in its several current manifestations, is one tool amongst many, a truly powerful tool with an ever expanding set of applications.[iv] The most valuable firms in the world are increasingly built with the whole toolbox, not just one tool (McGee and Chazan, 2020; Kenney, 2020).

The focus of this essay, nonetheless, is governance of the current iteration of the array of digital tools loosely labeled AI. Discussion about AI currently focuses on deep learning and machine learning (ML), but indeed deep learning and AI are simply the latest tools, labeled AI, in the toolbox.

There are, in our view, two separate debates about AI:

- Community and social preferences: One debate, about matters such as privacy and discrimination in applications, is really about molding AI usage to our community and social preferences. The array of discussions about AI ethics certainly fall here.
- Economic and strategic advantage: A second debate is whether, and for whom, AI creates economic and strategic advantage and how best to promote the development and competitive deployment of AI for economic and indeed geo-strategic, as well as military, advantage.

Quite obviously these two discussions often collide.

A third matter, which we address in the conclusion, is the question of how in an interconnected global economy we develop policy that carries significance in each national community.

Let us state our perspective clearly. The governance challenges in this era are not about an artificial general intelligence (AGI) that will make people into pets (Koebler, 2015). The threat is not the "Terminator" with human like capabilities, but rather about specific tools such as "Clearview" for facial recognition and "autonomous weapons" (Hill, 2020). The data-based, computation-intensive tools labeled AI are not going to wake up and pursue nefarious objectives of their own volition. The crucial governance questions are about how we deploy the tools—to what *ends*? with what *benefits,* and *to whom*? and with what *risks*?

Our purpose here is to highlight some of the questions and issues that governance must address, rather than provide a compelling approach to policy. *Part I* of this essay, *A Framework for Discussing AI*, sets a framework for discussing AI and focuses on what contemporary AI is and is not. *Part II, AI in the Community,* considers the interplay between the evolving AI tools and

our community norms. *Part III, Competition and Development*, considers AI's place in economic competition and development, and the effects of AI on work and workers. *Part IV, Turning the Governance Story Around*, considers AI as a tool of social and political governance. In *Part V,* by way of conclusion, we examine the international implications of the debates about governing AI.

Part 1: A framework for discussing AI: What it is and what it is not

The fourscore-year history of AI and its many definitions preface so many articles, trade books, and documentaries these days that we indulge only a few succinct comments on the history. As programmable computing machines emerged in the 1940's, British logician Alan Turing contemplated "intelligent machinery that could learn from experience … by altering its own instructions." In 1956, the discipline was born and named AI at the "Dartmouth Summer Research Project on Artificial Intelligence" (Leslie, 2019; Russel, 2019).[v] It has since experienced surges of public interest and funding (ca. 1956-1974, ca. 1981-1992), each followed by a so-called "AI winter" when funding and public interest were scarce. Around 2015 a new AI boom took hold and ".ai" began replacing ".com" on buildings and billboards.

The field has been defined by two parallel aspirations:

- A moonshot aspiration, "general AI" or "AGI", to build human-level intelligence in a computer program.
- Smaller, more or less well-defined aspirations, often characterized as AI's moving target ("AI is whatever we haven't solved yet"). These goals have included playing chess, transcribing speech, recognizing objects in a picture, diagnosing illnesses from a set of symptoms, answering questions based on a given text, proving theorems, and detecting nuclear tests in massive seismology data. These narrower aspirations are called "narrow AI".

Technologies developed by AI researchers had become commonplace in many systems but were not called AI even in the mid-2000's. Examples include speech recognition, image processing, data mining, industrial robotics, medical diagnosis, search engines, and recommendation systems for news, books and films. Their performance varied and, although highly valuable commercially, most performed below human-level.

As in all articles, we round the bend to discuss the power of deep learning. Around 2010, deep neural networks and related machine learning techniques began showing extraordinary results, many that approach or exceed human-level performance in areas such as speech transcription, face recognition, and medical image diagnostics. The anthropomorphic aspect of AGI seemed to come within reach; ignited a global AI "arms race;" drove Microsoft, Google, IBM, and many others to become "AI-first" companies; and seeded a wide range of efforts intended to assure that this newly powerful AI is safe and beneficial, and, certainly, to reassure customers and the public.

What AI is not (or at least not yet)

Personal digital assistants Alexa and Siri (named for marketing effect) respond sensibly to many spoken requests; OpenAI's GPT3 "writes" pages of coherent text in a consistent style based solely on a two-line prompt; Google Lens identifies plants and architectural landmarks from a single photo; and Waymo's autonomous cars safely navigate difficult construction zones on the streets of San Francisco. How far off is technology that can think, in general, as well as a human?

There are two problems with this leap. The first lies in the words, "in general." While each astonishing achievement of the last 10 years shows the power of the new AI clockwork, each is still a tool—and thus , narrow AI. *This is the essential of AI's importance these last 10 years.* The second problem is that, in colloquial use, the term AI implies the powerful aspiration to create a simulated human with the full spectrum of self-awareness, context-appropriate behavior, emotion, empathy, and the common sense about the world exhibited by a typical toddler or house pet.

AI academicians do not strive to create the conscious machines of Hollywood and literature. Machine uprisings, the "singularity," Iron Man's consort Jarvis, and uploading one's personal essence to the cloud all appeal to our draw to complex clockwork that comes alive, and to eternal life.

While exciting, AI is emphatically not machines opening their eyes and coming to life. The anthropomorphic implications of AI turn out not to be useful for today's discussion of impacts and governance. While the creation of conscious machines would have huge implications, it's fanciful to legislate about something that does not exist yet.

AI as a system of statistical inference

What, then, is AI? While, as noted above, it has no official, standard definition, AI tends to refer to technology that exhibits humanlike behavior. As noted, the defining behaviors and associated technologies have shifted over the decades. In today's most common usage, AI refers to systems that use ML, a sub-discipline of AI, trained on large corpora of data to make predictions and determine actions.

ML algorithms are engines of statistical inference, and thus come with the implicit limitations of such inference systems. Like all algorithms, they produce outputs in response to inputs from human users (clicks), sensors (images, temperature, position, acceleration), and accumulated data that has already been processed in some way. Where ordinary algorithms are hand-crafted to produce certain responses to given input states, ML algorithms "learn" which responses are best for a given input by iteratively adjusting many—often millions—of parameters based on large sets—often billions—of inputs. This is the "deep" in deep networks: there are multiple layers, each with millions of data points like image pixels, and the layers are interconnected by links that implement the "magic" through something called a loss function. In the case of "supervised learning," inputs are paired with desired outputs by adjusting the parameters for each link through a process called "back propagation," which requires a great deal of processing for large training data sets. In the case of "unsupervised learning," the system seeks clusters (e.g. "people who buy high-quality bird seed tend to like Harry Potter"). In "reinforcement learning," there is an element of exploration, such as moving pieces on a game board, and a reward function such as the game score, that enables a system to learn by experience. Then, in operation, the system calculates the most likely or best output for a given set of inputs based on this "training."

In today's ML focused AI, the very nature of statistical inference at the core of deep networks confers limits. As mere reflections of the correlations between inputs and outputs in data, today's most well-known AI systems are inherently conservative; they operate poorly outside the space of the data they have already seen. An AI system can transcribe a person's speech because it has seen the correct strings of words paired with the same or very similar vocal sounds; a different AI system can identify a dog in a photograph whose arrangement of pixels contain a canine because it has seen many similar pixel arrangements classified as "dog" in the training data. This narrows the domains where AI is effective, and the objectives it can feasibly achieve: training data is essential. Inputs can "fall between the cracks" of training data, yielding nonsensical responses. A data-trained tumor-detection system can diagnose more accurately than oncologists, but cannot have a hunch about a peculiar shape, e.g., that it reflects some extenuating circumstances of the patient. Other limits stem from the systems' superficial nature, as it is unable to draw common-sense conclusions based on modeling physics, causality, or human norms and preferences.

The astonishing breakthroughs in AI over the last decade have been driven by two things: (1) the explosion of available *data* in a growing set of domains from mobile and connected devices with sensors exchanging data on global, fast data networks, and (2) the drastic expansion of available computer processing power at ever lower prices predicted by Moore's Law.

What can today's AI systems do?

In addition to all of the techniques developed by AI researchers and successfully commercialized prior to the deep learning boom of the last 10 years, there are many new applications of deep learning, deep reinforcement learning, and variants, that have a wide range of capabilities with diverse and significant applications. Within these narrow domains, AI systems work with an immense speed and precision that no human can match. For example, AI is extremely adept at perceptual classification; AI systems can put names to faces and recognize common objects as fast or faster than most humans (Eckersley and Nasser, 2017).AI systems can likewise transcribe voices, translate between languages, locate tumors in diagnostic images, simulate styles, and maximize clicks on content. Famously, with millions of hours of simulated practice, AI can champion some of our most complex games. AI is unmatched in these and other domains that require discovering and reproducing patterns from extremely large amounts of data.

What are the risks associated with these capabilities?

AI can be used maliciously, and these uses warrant attention and efforts at mitigation. These uses include autonomous weapons, precision propaganda, and increasingly sophisticated cyberattacks, among others (Brundage et al., 2018). But equally important risks stem from the unintended consequences of AI systems deployed by actors with non-malicious intent. While the objectives of today's AI must be narrow, the impact of a system can spread far beyond its intended domain, particularly when systems achieve wide scale, as many have.

One example is the AI systems employed by social networks, whose effectiveness in keeping users interested can end up changing what they are interested in. By providing consistent and

attentive viewership from users, YouTube is highly appealing to advertisers. As such, YouTube's AI systems are tuned to the narrow objective of keeping users interested in each successive video in their automated queue. But as Professor Zeynep Tufekci of the University of North Carolina has observed, one of the algorithms' primary strategies for keeping users engaged is to show them ever more extreme versions of similar content (Friedersdorf, 2018). This can be as innocuous as showing videos about running ultramarathons after videos about jogging; or as potentially dangerous as videos of Donald Trump rallies lead to white supremacist rants and Holocaust denials. Performed at YouTube's immense scale, this strategy of engagement could alter how people think. A powerful statistical algorithm with a narrow objective has, thus, broad social consequences that by definition could not be incorporated into its reward model.

Another example of AI's unintended consequences beyond the fulfillment of narrow objectives lies in the facial recognition company ClearView. With a library of over three billion images, ClearView empowers customers to upload a photo of an individual and in response learn the identity and any discoverable information on the internet about that person (Hill, 2020). A New York Times investigation in 2019 found that ClearView had been adopted by over 600 law enforcement agencies worldwide, with little to no scrutiny. With its massive library of data, its algorithm is remarkably effective at the helping security services identify individuals. But the very scale underlying the system's effectiveness has resulted in a significant impact on the much broader domain of personal privacy.

Broader unintended consequences are inherent to the combination of narrow objectives and significant scale. Efficiency is a limited metric when evaluating the pursuit of goals by powerful entities; but efficiency is the narrow objective that most of today's systems are optimized to achieve. This makes the risks of today's AI difficult to mitigate, and suggests that greater caution is needed in deploying systems at scale.

What are the limitations of today's AI systems?[vi]

Beyond their risks, AI's recent achievements have also led to misconceptions about its capabilities. Articles have been published that variously claim AI can read, drive cars, understand emotions, and create art and music, among many other abilities that have long been the exclusive purview of human beings (Marr, 2019).These claims are understandably exciting for journalists, but wrong to computer scientists. AI systems may be able to answer questions about a book after analyzing its pages; but these systems cleverly match text in the questions to corresponding spans of text in the book, processed in such a way as to allow for variation (Marcus and Davis, 2019; Joshi et al., 2017)). This is nothing like the synthesis of new knowledge with old that constitutes reading for understanding; an AI system could never build a rocket after analyzing a book on rocketry. Autonomous vehicle systems can detect objects in their path; but they cannot see that a bale of hay is about to fall off the back of a truck, or that if a ball suddenly rolls into the street, a child may well follow it (Knight, 2020).Marketers will find it useful that AI systems can pair animated facial expressions with broad categories of emotional states, but this is not the same as exhibiting empathy or connection. And art-generating AI systems are striking in their ability to copy styles, but random at creating substance. For example, the accomplished "computer artist" Harold Cohen (1928-2016) viewed AARON, the AI program he wrote and maintained through his career, as his collaborator, "at times responsible

for the composition, coloring and other aspects of a work" (*Harold Cohn, artists – obituary,* 2016). However, its creations depended heavily on randomization (Deutsch, 2016).

Some of AI's limits can be broadly indicated by the inability to generalize. Human children, as has been argued by Professor Allison Gopnik of UC Berkeley, develop general frames or models of the world, shuffling amongst models to learn and interpret what is going on around them (Gopnik, 2011; Samuel, 2020). AI systems have no such frames or models to refer to. As Gary Markus has written, AI systems do not know what they are for; nothing in their necessarily narrow data tells them anything about the purpose of their task, which depends on the world in which it sits (Marcus and Davis, 2019).

The inability to generalize prevents current AI systems from three forms of reasoning, which power humans' interaction with the world. First, AI systems cannot yet reason about causality. By observing daily life around them, humans at a very young age understand how a physical action causes a result. But as has been documented by MIT Professor Joshua Tennenbaum, AI systems cannot answer basic causal questions about a scene, such as "what caused the ball to collide with the cube?" (Knight, 2020) A machine may detect a pattern in image pixels that represents a ball colliding with a cube, but has no general model of the world telling it that the ball moved because a person pushed it. This has obvious implications for AI systems intended to function in a physical environment, as with the autonomous vehicle systems noted above. But a baby's ability to reason causally about the physical world underlies their ability to reason about cause and effect in the abstract as they get older. This, in turn, is a core factor in human's ability to be effective in situations that have not been encountered before (Knight, 2020). So long as AI's reasoning is confined to detecting correlations among arrangements of pixels, words, or other superficial data, it will not be able to work through unfamiliar situations.

Second, AI systems cannot yet understand human emotion. Human emotional understanding is rooted in an ability to know that one feels what another feels, and in a shared grounding in the human condition. Neurons indicating the ability to signal "I feel what you feel", to empathize, have been observed in social animals beyond humans, including chimpanzees (Blakeslee, 2006). But such interactions can only be simulated in AI systems, which have no background understanding or sense of shared experience. Additionally, simulated understanding of emotion is in its infancy in AI research.

Third, AI systems are unable to use judgment. Judgment is the ability to step back from the immediate domain and understand something's significance to the broader world., Humans exercise judgment by relating an immediate event to their general model of the world and reason about how they are related to one another. But while today's AI systems may be able to detect something unusual in a pattern, they lack the general world model necessary to determine whether the appropriate response is "that's odd," or "eureka!"

What is going on here? Interpreting reality

Humans' distinctive aptitude for these three types of reason (identifying causation, understanding emotion, and exercising judgement and generalizing), can perhaps be assessed in terms of three core aspects of our cognition and interpretation: *context*, *narrative*, and *worldview*.

Context refers to how we answer the question, "what is going on here?"[vii] Mr. Robot certainly does not know.

In some sense, *context*—derived from our "models" of how the world works and our frames of reference to the world—, defines for us the answers to the question, "what is going on here." In characterizing context, we are specifying which elements in a situation are relevant and which elements fall to the background. Of course, the aspects relevant to a context are continuously changing. For example, in high school, you and friends are being boisterous and disruptive before class. Then an adult walks in: is that person a teacher who demands attention, or a custodian who can be noticed and ignored? The answer to this question rests in how we define context, which constitutes our frames of reference for how the world works.

Seen thus, specification of context is a uniquely human capacity, linked to our experiences and the way we understand the world. That conclusion forces a question: in which applications can we apply AI tools without a human definition of context? In other words, when is a narrower specification of context sufficient? In which applications is a human definition of context essential?

Contexts do not stand alone. They are part of a sequence that constitute the *narratives* of our lives. Defining context involves defining what "story" one is in at any given moment—what is the narrative? We live in socially defined narratives with the people around us, the rewards and threats we face, and much more. Different narratives imply different definitions of context; if how one defines context depends on his or her "narrative" about life, then, in similar situations, different people will define contexts in different ways.

Our narratives are themselves framed within, what is labeled in social science, "*worldviews.*"[viii] Narratives about ourselves, others, and the world can be built into broader narratives about our communities and others, that constitute a view of how the world works as a whole. Religion, economic philosophy, or any number of broader narratives, world views, create the environment in which our individual narratives emerge.

Taken together, context, narrative, and worldview comprise much of what is distinctive about human intelligence, and as yet beyond the reach of statistics-based AI systems. Our continuous definition and redefinition of context, and ability to refer back and forth between it and the immediate objective, is essential to our ability to understand causality, emotion, and judgment and make essential generalizations.

As generalizability is a core aspect of AGI this casts doubt on the possibility that systems built on today's dominant ML models can achieve AGI. To be sure, significant breakthroughs are needed. The recent abatement in significant AI breakthroughs suggests that AGI is a long way off.

Part 2: AI in the Community

To begin to clarify issues and choices in the governance of AI, we focus on community. The diverse set of applications of AI tools or techniques, narrow AI in the parlance, forces a

debate about an entire panorama of community norms and public policies. The list of AI applications, and the issues they raise, is endless. It runs from criminal justice (sentencing and resource deployment) to surveillance (facial recognition and monitoring of email traffic) to medical applications (drug development strategies and treatment recommendations) to retail purchases with platforms like Amazon and, indeed, social discourse and news feeds. For some, risks, such as autonomous weapons or badly designed FinTech tools that suddenly undermine the financial system, are concerns. We set aside those dramatic risks, including the potential emergence of an AGI that may pose risks at a societal scale or worse, though such concerns are part of Mark Nitzberg's (co-author) professional responsibilities and commitments. We focus, rather, on AI in the community. Applications raise interwoven concerns that go well beyond privacy, an issue which EU and California legislation have usefully begun to address, to include bias in hiring as well as in criminal justice, for example. Surveillance of our activities; targeting of news and advertisements as Facebook, simply an example, alters social and political discourse.

Applying existing norms and rules to an era of AI, as difficult as that would be, does not properly define the governance challenge (Lessig, 2000; Lessig, 1999). Existing norms and rules express social practice as well as explicit political deals. Social practice evolves; those deals are likely to have to be remade, and entirely new issues requiring new norms and rules emerge. As an example of emerging norms, note the reconsideration of criminal justice with the Black Lives Matter (BLM) movement. Decision processes where AI based tools are used forces a look at historical data, and that historical data embeds earlier deals and norms. The history embeds bias, bias expressed in the data. In general, the reality that will emerge from debates about AI will depend on the political and social context in which the issues present themselves.

One might ask, to reverse the logic, whether AI tools can force us to confront sources of bias. Judicial systems and policing, as the current upheavals around BLM make clear, are systematically biased. The video functions of smartphones have enabled public awareness of police violence against the Black community and other communities of color. While AI tools do not create past bias, the bias is embedded in the training data, they do automate processes and decisions based on past behavior, they can make bias more evident. Perhaps AI tools, though less publicly accessible than smartphones, can have a similar consequence. Of course, this is debatable; as Cathy O'Neil would argue, algorithms embed human biases into concrete and seemingly irrefutable outputs, and these algorithms are scalable and self-reinforcing. The outputs are highly consequential for individuals, but the process of arriving at these outcomes is nearly impossible to understand without knowing the variables in the algorithm or being a highly skilled computer scientist (O'Neil, 2016; Silva and Kenney, 2019).[ix] Consequently, algorithmic transparency may not have much meat. This is a point we return to later in the essay.

Assuming that a community could settle the objectives they want to pursue in governing AI, the questions remain: who should do it and what should they do?[x]

Framing the governance problem: who governs AI and what should they govern?

Who should govern AI? With the creation of an Organization for Economic Co-operation and Development AI Policy Observatory and the proposal for a United Nations Panel on

Artificial Intelligence, there is a swirling debate about how to govern AI.[xi] This is not simply a matter of abstracted ethics, but of which public institutions and whose institutions are best situated for regulation and governance. Should the regulation be at the city level as has often occurred with Uber? At the state or regional level as with California legislation on privacy, which was inspired by EU privacy law? At the international level, as bargains amongst states?

Indeed, what should be undertaken by public authorities, and what by private self-regulation? Individual instances of calamities emerging from private self-regulation, such as the Boeing 737 Max, underpin the importance of balancing public authorities and private actors in regulation. Google's AI Principles are another example of private self-regulation; are the principles merely suggestions or does Google have a system for accountability? (Google, 2018) More generally, do public authorities require the detailed technical know-how embedded in private sector operations? If so, what should be the balance between private actors and public authorities as they collaborate on product regulation? Digital platforms represent the systemic problem of AI governance. Together, the algorithmic structure of the platforms and the "terms and conditions" to which those engaged on the platform agree make the digital platforms private regulators that are difficult to control by public authorities (Cutolo and Kenney, 2020; Kenney and Zysman, 2016).

This matter of where debates about AI are held and the fora in which AI rules are set is not a technical matter. Different interests are represented in different fora. As a result, different outcomes can be expected in different fora. *The decision of where and how to frame the debate over AI governance will influence, if not determine, the results.*

What should be governed? Even if we settle the matter of who governs, we are left with the challenge of deciding what in fact should be governed to govern AI. This is more complicated than a simple statement suggests.

To start, we ask whether we ought to focus on the tool itself, AI, or the applications in the domains to which that tool is applied. Given the diversity of domains, focusing on the tool itself risks elevating the debate to abstractions that are ultimately empty. That can make it difficult to address specific challenges. The mantra in AI conferences for several years, FAT (Fairness, Accountability, Transparency—or, adding Ethics, "FATE") may be easy to agree on in principle, but difficult to operationally define or implement.

We need, it would seem, to consider particular domains and applications to move from the abstract and general. The same AI application in different sectors will generate distinct gains, consequences, and risks. The uses of AI that may be problematic in one domain may not have the same significance in another. A mistaken advertisement for shirts that draws the reader into a cycle of shirt advertisements is of less significance than targeted political misinformation or mistakes about medical diagnoses.

Not all crucial questions, however, can be identified or addressed in specific, narrow, contexts. Issues that reappear across sectors may be hidden or underestimated because the issue is not central in a particular domain. Some issues, such as bias, may be most evident or crucial in some domains, and most easily identified there, but are important throughout society.[xii] Many issues

that run across applications might have common solutions such as data privacy and surveillance. A general overview will still be needed, often stepping beyond the AI tools themselves.

Let us repose the question. What elements of AI tools or applications should be our focus? There will be debates about which aspects of AI tools should be the focus of governance. A focus on machine learning and deep learning algorithms that are at the core of contemporary AI will certainly not be sufficient. Phrased in our language, that would be a focus on algorithms, data, actors, and outcomes.

Algorithms often seem to be the magic wand. While the AI software, the algorithms, may be thought of as the magic wand that when waved over data produces magical outcomes, the algorithms themselves are hard to review, regulate, and govern for a number of reasons. Transparency, one demand in debates about AI, would be at best difficult.

Unpacking the software code and understanding, in isolation, the consequences of a particular algorithm's code would be a remarkable challenge in each instance, requiring quite skilled and trained technical staff. In some cases involving deep learning algorithms, it may not be logically possible to unpack and explain the processes that produce a particular outcome. As AI becomes as embedded in the economy as electrical wire, we may need standards for its use, but reviewing each algorithm is not an effective approach to governance.

However, if we cannot unpack the code, can we understand the logical architecture of the algorithm, what is being considered, how the elements are related to each other, and how they are weighted? Assume for a moment that it is possible, and possible at a reasonable cost, to unpack and examine the logical structure of AI algorithms. While that is certainly debatable, two important questions still arise:

- First, the architecture and logic are part of intellectual property. Under what circumstances can a regulatory agency insist on access? And which agencies should have the privilege? Should an EU or Chinese authority have the right to look at the architecture of a Google or Facebook algorithm? Or should a German authority have the right to look at a Tesla algorithm? Undoubtedly there would be endless legal challenges to such access.
- Second, what circumstances would justify starting down this path to transparency?
    - Data, how it is gathered and deployed, by whom, and to what ends underpins how these tools can be used. Can we approach AI governance without looking at the raw materials on which it depends? That opens directly to issues of data sharing and ownership, as well as of data privacy.
    - Online two-sided platforms, digital platforms such as Google, Amazon, and Facebook amass data and apply these AI tools in diverse activities. Addressing the AI questions will inevitably force us to consider issues of platform regulation.

So, let us turn our discussion to data**.**

Data, certainly, fuels the game. Can we approach AI governance without looking at the raw materials on which it depends? Therefore, a focus complementary to algorithms, is data. Data,

how it is gathered and deployed, by whom, and to what ends underpins how these tools can be used. The capacity to gather, store, manage, and analyze large amounts of data is what is new.

At its core, of course, algorithms—machine learning and deep learning—are statistical inference engines, attempting to predict the future based on an examination of past behavior. Bias, then, may be built into the logic of the algorithm, what the logical analysis sets up. However, perhaps even more likely, the data that is fed into the analysis has history of judgments that may, or may not, be valid. Since AI tools are statistical inference engines that simply work with existing biases in the data or the design of the undertaking, biases built into the existing data will be reflected in outcomes. And, without delving too deeply into logics of social behavior, predictions often generate outcomes. For instance, tell a teacher that based on a formal test a particular student is very talented and likely to do better in the years going forward than in past years, and an uptick in student performance will result. Unfortunately, predictions may reinforce past patterns of bias. There are feedback loops (O'Neil, 2016).

Regulating AI, thus, overlaps with regulation of data. The AI tools run on what is now loosely called big data. So, the crucial questions asked more generally about data and data privacy take on increased meaning in an examination of AI. We must not be drawn here into a general discussion of data regulation and governance. From an AI vantage the concerns include:

- Who collects what data?
- Who can use the data and for what purposes?
    - How is the data shared?
    - How is use of the data controlled or priced?
    - Is the quality of the data, biases built in the data sets themselves, vetted for the users?

Some data issues, of course, vary by domain. Consider that in retail, knowing customer buying patterns is an asset that may be protected for use by one firm, or that firm may sell access to customer data to another firm. The issue, as a result, is whether consumers should be compensated for their data: who can use it? How can consumer privacy be preserved? By contrast, in many business-to-business applications, for example, in a production system using the tools of multiple vendors, aspects of the data must be shared to make the system work and other aspects will need to remain proprietary as part of firm and tool specific intellectual property. Governing AI does not stand alone. Crucially, governing AI simply opens out into the core issues of the digital era, of the information society.

We must also consider the *actors themselves,* the intent and purposes to which they put AI tools. That ultimately means governing the actors. Most evidently, we might consider regulating digital platforms as a means of regulating AI uses. Two-sided, digital platforms such as Google, Amazon, Facebook, Uber, and Yelp, amass data and apply these AI tools in diverse activities. Addressing the issues of AI applications in platforms will inevitably force us to consider issues of platform regulation.

Outcomes are really the issue. The obvious question, forced by an effort to govern AI, is what are the norms we want to support and encourage and the behaviors to be discouraged or banned? Indeed, the answers, outcomes of political conflict and social debate, do not simply pre-exist,

awaiting implementation. Rather, the values, objectives, and preferred outcomes will be created by the process of creating a governance system.

If we focus on outcomes, the obvious issue is whether to consider general rules or domain specific concerns. As noted before, the types of data or tools that may be acceptable to make retail offers, or the scrutiny about those tools, may differ radically from the types of offers and scrutiny appropriate with financial offers or medical recommendations. Indeed, that is already the case: the Securities and Exchange Commission may review the financial offerings of Macy's but not their sales. So many of the issues must be debated and rules implemented, domain by domain. Moreover, some actors, such as digital platforms, cut across domains, and may require actor specific regulation of their deployment of AI tools. In some sense, it is a matter of how we want to derive general rules and figure out how to apply them across domains or how we want to infer general rules from sectoral domain situations. Certainly, there will be general principles, such as avoiding biased outcomes or limiting the extent of surveillance, that should frame particular domains or specific actor regulations.

In sum, AI raises new challenges and issues. Rules about AI cannot simply be grafted on pre-existing norms and regulations.

Part 3. AI, Competition, Growth, and Jobs

The implications of AI for economic growth and employment as well as for geopolitical competition is, for many, the central concern. Indeed, these concerns are, arguably, the drivers of public investment in AI.

Investing in AI for growth and power

Ought countries be investing massively to promote leadership in AI technology? Indeed, what would that mean? Certainly, a case can be made that effective development and deployment of digital technology— now primarily focused on AI—is essential for a position of at least credible deterrence and influence in cyber conflicts affecting national security (Lee, 2018). In any case, it is debatable whether basic AI research spills directly into commercial applications, and often not easily, or at least not automatically into deployable national security application. National security concerns, though, are complicated issues both in basic science collaboration and the development of specific applications. Let us set aside this discussion for another day, or essay.

Our question here is how to promote AI potential for economic development.[xiii] One vantage is that effective economic exploitation of AI possibilities is principally a question of leadership in basic technology and computer science. If so, heavy investment in advanced research, say at University of California, Berkeley or Technical University of Munich in Germany, and very advanced projects will be called for. The difficulty is that the results of investments in basic research in public institutions in the West quickly spread internationally and are very difficult, often impossible, to hoard for a particular national research agenda. Basic science advances internationally, unless sharp constraints are imposed on the flow of information, and even then, such flow often continues. The resulting question would be what international research alliances

are called for and what will be the terms of sharing in the development, not just after the fact access to results.

An alternate view is that harvesting and deploying cutting edge research is central. Certainly, a nation's, or a region's, research is essential if domestic, or local, institutions are to participate in the global flow of research. In this view, the development and specific application of widely available AI technological foundations is central. Indeed, in this case, the proper policy responses are likely to be found in discussion of specific sectors, as we argued is the case with issues involving social and community norms, and in programs to diffuse capacities to deploy AI possibilities.

In any case, policy promoting the deployment of basic AI knowledge into applications through society and the economy raise significant challenges:

- **Business strategy**: How does one encourage businesses to identify, let alone adapt, new applications? Will only the largest, best funded firms, or venture capital subsidized start-ups be able to afford the process of discovery and experimentation that deployment requires? Some sectors may be winner-take-all, but throughout the economy, being a fast follower may be more effective (Andrews, Criscuolo, and Gal, 2016). What is most effective as public policy with the case of AI?
- **Workforce development and training**: Wide deployment will require a broadly trained workforce capable of using these tools or the development of user interfaces that make application simple. In any case, an AI for growth strategy calls for investments in labor, workforces, skills, and software to facilitate human-computer complementarity. Training subsidies, direct or through tax breaks, might be called for. In the US, that might also mean investment in community colleges, certificate programs, and other short-term high-value degrees, as well as in secondary education, and, in Germany, adapting apprenticeship programs in real-time.
- **Secure trusted applications**: Moreover, if the challenge is deployment of AI based tools for economic growth, then the diffusion, perhaps a requirement of effective cyber security, may be essential to assure trust in the integrity of those tools, which will permit widespread use of trusted applications.[xiv]

The answer of how much and how to invest in advanced research, and how much and how to invest in specific applications and deployment will shape the level and mechanisms of government programs. Cynically perhaps, since no indisputable answer is possible, the political and administrative winners within the government will set the balance of investments. What, then, will be the consequences of the development and deployment of AI? Will it accelerate growth? How will it affect the workforce?

Will AI accelerate growth?

Assume policy is successful in encouraging AI development and deployment in the economy. Will AI accelerate growth? We need to proceed carefully. The box , the suite, of intelligent tools, AI being one, will certainly bring dramatic changes in goods and services, in processes of creation and production, and of distribution. That said, we need to be very careful

about the claims that this might unleash unprecedented growth. Consulting firms, whom one might suggest were promoting their own services to potential clients, suggest that in 10 to 15 years AI could double growth rates and add 10+ trillions, dollars or Euros, to global GDP (Funk, 2020). The easy consulting firm conclusion is that leaders—companies and countries—will capture the benefits. The implication is that laggards will be disadvantaged. Cynically, the implication is that their consulting services can help the client firm or country to be a leader.[xv] Those discussions, moreover, tend to ignore altogether the distributional questions. [xvi]

There are rather basic questions, even setting aside broader debates about the sources of economic growth.

First, how, precisely, will AI tools increase the output of what the whole economy generates from existing capital and labor? Based on historical evidence, traditional economic theory would predict that new technologies lead to increases in productivity, which increase incomes and thereby increase the demand for goods and services. This increases the demand for labor and output in certain occupations, industries, and sectors. Will AI tools just create advantage for the leaders in particular sectors or domains that are best positioned to deploy the tools? In that sense, is the deployment of AI any different from a sequence of basic technologies over the years?

Second, fundamental technologies (e.g. railroads) depend on related suites of technologies (e.g., metal for rail and engines) and often on underlying base technologies (e.g. steam). If we invest in AI without the rest, what will we gain? In that sense, digital platforms are, in the context, an essential part of the suite of technologies.[xvii] The impact of AI tools depends on the development and deployment of the suite of technologies.

Third, as part of that suite of tools, appropriate infrastructure is required and needs to be singled out. For AI to have an economy-wide impact requires a reconsideration of the available and needed infrastructure and policies. For instance, national markets required the railroad. Trucking innovation required interstate highways. Widespread access to high-speed internet is needed for digital innovation, and clearly, AI applications.

Fourth, and finally for this discussion, there are an array of obstacles to even beginning, let alone successfully completing, development and deployment. For a company to capture the full benefit of AI development and deployment, fundamental rethinking of business strategies, corporate structure and responsibility, and work organization will be required. All of this influences the kinds of skills required in the workforce itself. Rethinking business strategies is hard enough; but that often requires existing leadership to go in new directions, which is, shall we say, difficult for established management. This is indeed the much-acclaimed innovator's dilemma (Christensen, 1997). Is a semi-conductor firm selling chips or systems? Similarly, as John Deere moves from selling tractors to farm management services, different leadership skills will likely be required.[xviii] The effective implementation of AI tools will likely move across a generation. The impact on growth is even less certain.

AI, distribution, and the impact on the workforce

One element of the AI and growth story, the impact of AI-enabled automation on the workforce, is of particular concern to policy makers. Assume that firms effectively deploy intelligent tools, with AI firmly rooted in them, to win in the marketplace. What will happen to the workforce? The simple conclusion in this essay about governance is that the consequences of deployment for the workforce hinges on policy choices, not the technology per se. [xix]

First, sorting through the diverse claims about the impact of AI on jobs, on employment, is as difficult and problematic as sorting through the claims about growth. There have been competing claims that run from imminent disaster to slowly evolving disaster to perfectly manageable traditional reorganization of economy. The extensive and diverse body of research[xx] focuses variously on tasks that may be displaced, on jobs lost, and on sectors that may be changed. For the most part, the research does *not* focus on how these tools affect the organization, or better, the reorganization of work and responsibility, of how the activities, generally framed, of production, client or sales, or management can be reimagined and reconceived. Simplistically, we all know that jobs will be created, destroyed, modified, and transformed. What we don't know is how fast or what that transformation of work will look like. Indeed, there is a dispute about whether the deployment of these tools will down-skill the core of the economy, a dominant view at the moment, or create new possibilities and opportunities (Turea, 2016). There is a consensus view among economists on many aspects of the labor market issues: there is no evidence of long-term technological unemployment, productivity and employment growth go together, and the dislocation effects associated with changes in the demand for labor fall unevenly across workers, communities, occupations, and sectors (Tyson, 2019). In our view the consequences will likely depend on the deployment strategies adopted by firms and how firms understand the market advantages these tools can create.[xxi] In that sense, the labor market issues are wide open.[xxii]

Part of an AI governance strategy should be to address the employment question. Encouraging the trajectory of the deployment of intelligent tools, all the more powerful with AI applications, to support the upgrading of existing work and the creation of new work is essential, and possible.[xxiii]

In sum, government promotion of AI, a crucial part of governance, has at least two dimensions. One dimension is to encourage the development, diffusion, and use of the technologies. Setting aside purely military applications, achieving that goal requires assuring societal capacity to absorb and diffuse cutting-edge technology. The underlying science, military technology aside, will likely be widely available. Diffusion and application is essential, and that is a very different problem than investing in basic research or even development-oriented research. The diffusion question is rarely fully addressed.

The second dimension is the impact of AI on growth and employment. Strategies to capture the *growth possibilities* of intelligent tools in general, and AI in particular, require complementary development of a suite of complementary technologies and investment in a skilled workforce. Part of that is to promote business and deployment strategies that encourage complementarity and the upgrading of work and work opportunities.

Part 4. Turning the Governance Story Around: AI in Government and Policy

Application of AI tools to governance and government functions may seem superficially attractive. One may argue that such tools work in retail, political campaigns, industrial maintenance, and operations. So why not use these tools to make government more effective? Before rushing to a conclusion, there are two matters that must be considered.

First, the argument that data and applications and outcomes need to be carefully vetted is all the more important in governmental functions. Biases built into data, which may be significant in evaluating private recommendations and outcomes, are all the more important in the supposedly even-handed handling of public affairs. Vetting how data is constructed and taking seriously, and reviewing, concerns about distribution of recommendations and outcomes, will be essential to assuring transparency and building trust in AI supported public sector applications.

There is a *second*, and just as significant, concern. Consider that the notion of "smart cities" suggests that urban networked functions, such as transportation, governed by digital intelligence can improve urban life. Perhaps, using AI, available resources can be more efficiently deployed. Should we not make more efficient use of public resources? But let us dig deeper before rushing out the AI Force.

The pursuit of efficiency facilitated by AI can obfuscate, hide, disputes over what the purposes and goals of policy should be. Those disputes about purpose and goals are the politics of governance and usually even more important than narrowly defined efficiency.

Efficiency assumes a specification of goals. However, in politics and policy the goals themselves are in dispute. Many decades ago, a candidate for mayor of Grenoble France, an engineer, opined that there was a single best way of collecting garbage. That assumed that goals had been set and the distribution of garbage collection resources allocated. But, of course, that avoids questions such as whether all neighborhoods will have the same level of service; do all customers in all neighborhoods—residence and business—have the same needs? Are wealthy neighborhoods better served than less well-off communities? Indeed, the allocation of garbage collecting resources is a value judgment about the goals. Game theory, for example, only makes sense as an analytic tool if we know what the game is and can agree about the goals, the preferred outcomes. But what if the game itself or the outcomes favored in the game, are subject to an encompassing political and social game? Setting those goals, those values, is the very substance of policy and politics. The political winners, certainly, set the goals and often decide the game being played. Efficiency can only follow those decisions, and fights about efficiency are often, in fact, clouding fights about goals and values.

Some might respond that with very sophisticated survey tools, or observations of citizen choices, a bit like Amazon's understanding of retail commerce, analytic tools can help us assess a community's values, and more easily define objectives. Then, such an analysis would suggest we could move the game of politics into the AI tool set and revert to the matter of efficiently pursuing established goals.

Unfortunately for that line of reasoning, there is a deeper problem. Set aside the question of how we truly assess a community or polity's values at any given moment—itself a matter of debate.

Even if we could "measure" or "infer" the community preferences at a given moment, the problem is not resolved. Why not?

Social values are not fixed. They are not simply the sum of individual opinion at a given moment. Rather, social values are the product of social conflict, of conflicts amongst groups. Values evolve over time. Indeed, the political actors and indeed the social "groups" in a community are in constant reformulation. If we did not already know this, the recent powerful surge of the Black Lives Matter movement and the large shift in measured public opinion demonstrates this. In 2016, Colin Kaepernick, San Francisco quarterback, began taking a knee during the national anthem in support of social justice. He was then shunned by the NFL, the National Football League. By 2020, in the NBA, the National Basketball Association, those not taking a knee were the exception. The issues around which communities organize and about which social groups fight are constantly evolving, both as a product of social and economic circumstances and the logic of political competition. The context, the definition of a situation in which we search for goals and means of accomplishing those goals, is itself a part of the fight.

Consequently, a system of AI governance will reflect the values of those who design the system in the first place, perhaps entrenching their values and submerging the processes of social and value development. That will be true whether it is a transport system or a system of policing. The question of efficiency, the least cost achievement of agreed policy goals, or effectiveness, and how fully a goal is achieved, only arise when the goal is settled. And those goals will <u>always</u> be contested.

Part 5. Conclusion – Implementing Governance

By way of review of the issues identified here, let us consider the difficulties of translating policy preferences into practical policy and specific regulation. We reiterate that policy concerns will be with deployment of "narrow" AI tools and must begin with understanding the inherently conservative character of machine learning and deep learning, which are instruments of statistical inference building on prior data. There are some basic implications of our discussion.

First, we must recall that simply declaring objectives—be they digital privacy, transparency, or avoiding bias—is not sufficient. We must decide what the goals will be in operational terms. And equally difficult, those goals must then be translated into digital operations, recalling as always that code is its own form of law (Lessig, 2000; Lessig, 1999).

Second, AI applications are part of a suite of intelligent tools and systems that ultimately must be regulated as a set. Digital platforms, for example, generate the pools of big data on which AI tools operate and hence, the regulation of digital platforms and of big data is part of the challenge of governing AI. Many of the platform offerings are, in fact, deployments of AI tools. Hence, focusing on AI alone distorts the governance problem.

Third, the issues and choices will differ by sector. The consequences, for example, of bias and error will differ from a medical or a criminal justice domain to one of retail sales. Creating alternative mega-platforms for search or shopping may be quite difficult and breaking them up often pointless chimera. By contrast, alternative digital platforms for finance are being discussed

as part of consideration of central bank digital currency (Bank for International Settlements, 2020).

Fourth, the economic implications of AI applications are easily exaggerated. The possibilities for accelerating growth are overstated by would-be consultants, while the consequences for work depend less on the technology itself than how it is deployed in the reorganization of work in this era of intelligent tools and systems. Should public investment be concentrated on advancing basic research or on the diffusion of tools and the user interfaces and training needed to implement them?

Fifth, the application of AI tools in public policy decision making, in the design of transport or waste disposal or policing, or in a whole variety of domains, requires great care. There are substantial risks of confusing efficiency with public debate about what the goals should be in the first place. Indeed, public values evolve as part of social and political conflict.

As difficult as it will be to decide on goals and a strategy to implement the goals in one community, international agreement may be essential to actualize AI governance. That said, any agreement that goes beyond simple statements of hoped for outcomes is very unlikely. Let us consider why.

Even at the most basic level, the globally interconnected character of the digital systems means that a policy goal in one community will often open contentious international debate and conflict. To take a curmudgeonly stance, since we will not find common agreement on goals, international governance will be mostly talk. The *EU privacy initiatives* did spark a complementary legislative initiative in California. Perhaps eventually there might be a Transatlantic deal. But can we imagine an agreement including China on the use of facial recognition tools? *French tax* initiatives on digital platforms were rebuffed by the Trump administration and European *competition policy* objectives would probably require another Transatlantic deal to take real form. Lurking on the sidelines, for the moment, is a debate about foreign takeovers and State subsidies for digital firms. Even more contentious, as the Huawei struggle clearly shows, there are outright competing objectives in *cyber security*. Huawei was arguably a product of Chinese state subsidy and potentially, if not actually, an instrument of the Chinese state. Whatever the reality, a full-fledged Western response to the Huawei challenge will require a European/American producer of 5G gear, with the most likely candidates being Nokia and Ericsson.[xxiv] While we are arguing to block Chinese state subsidy of its digital firms, efforts are likely to achieve limited success at best, promoting a 5G producer as an alternative to Huawei will almost certainly require state subsidy, or rather subsidy in several forms by a number of states. Would an agreement on the use of *targeted social media* that was adhered to by China and Russia be possible?

Indeed, more likely than agreement on global governance, there will be intense cyber rivalries that risk splintering the digital world into two (West vs. China), or less likely three systems (US vs. Europe vs. China) linked but separately developing. It is not just that there will be conflict over the community values to be implemented in each, or about strategies for national development. Rather, there will certainly be a conflict over the development of AI tools with intelligence, security, and military significance, sparking investment in AI.

We note in conclusion that post World War II trade policy reflected bargains, not always on equal footing, between the US and Europe (Griffith, Steinberg, and Zysman, 2017). AI governance, for Europe and the US, we would argue, requires that once again we have a Transatlantic deal. Only such a deal about how to proceed with not only AI but the suite of intelligent tools, can shape the terms of the digital revolution, of which AI development and governance is a crucial part.

Please note that the footnotes are incomplete. This is a working draft for comment.

[i] Professor Emeritus, Political Science, University of California, Berkeley; Co-director, Berkeley Roundtable on the International Economy (BRIE); Convenor of the WITS Project at BRIE/CITRIS.

[ii] Center for Human-Compatible Artificial Intelligence (CHAI), Berkeley AI Research (BAIR), University of California, Berkeley.

[iii] Gary Markus (Markus, G. 2018. "Deep Learning: A Critical Appraisal." *arXiv.* https://arxiv.org/abs/1801.00631)wrote: "As a measure of progress, it is worth considering a somewhat pessimistic piece I wrote for *The New Yorker* five years ago, conjecturing that "deep learning is only part of the larger challenge of building intelligent machines" because "such techniques lack ways of representing causal relationships (such as between diseases and their symptoms), and are likely to face challenges in acquiring abstract ideas like "sibling" or "identical to." They have no obvious ways of performing logical inferences, and they are also still a long way from integrating abstract knowledge, such as information about what objects are, what they are for, and how they are typically used" (p. 23).

[iv] This position is argued by an array of folks: Mark Nitzberg, Ken Goldberg, and Gary Markus amongst others.

[v] D. Leslie, wrote in a review of Stuart Russel's *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking Press 2019: Interestingly, there was a debate at that conference about what name to apply and AI was chosen to attract attention. Herbert Simon and Allen Newell proposed complex information processing. John McCarthy and Marvin Minsky proposed AI for marketing and aesthetic reasons. See:Leslie, D. 2019. "Raging Robots and Hapless Human: The AI Dystopia." Nature 574: 32-33. https://doi.org/10.1038/d41586-019-02939-0.

[vi] The limitations of narrow AI, our focus in this essay, overlap with arguments about why AGI cannot be achieved. The arguments about AGI are often rooted in philosophic considerations of human intelligence. Without fully developing those issues, we would refer readers to these materials: Fjelland, 2020 and Mitchell, 2019. Fjelland, R. 2020. "Why General AI Will Not Be Realized." *Humanities and Social Science Communications* 7, no. 10: 1-9 https://doi.org/10.1057/s41599-020-0494-4

[vii] For us, the robot's inherent inability to understand what is going on recalls the confusion of Mr. Jones in Bob Dylan's "Ballad of a Thin Man."

[viii] The notion of world view is widely discussed in social science literature.

[ix] O'Neil, C. 2016. *Weapons of Math Destruction.* New York: Crown Books. O'Neil wrote "One difficulty with understanding and rectifying bias in algorithmic processes is that the location of the bias varies. In some cases, it is embedded in the training data and often is indirect rather than direct, i.e, it is not the data is deliberately biased, but rather when analyzed other often unexpected variables impart the bias. For example, the US population is highly segregated and thus using zip codes could create racial bias. Bias may also be the results of the algorithms that unconsciously create bias, rather than the algorithm being biased, the users of software or platform may be biased and the algorithm would "learn" bias from them and then feed that bias back into future decision-making."

For a model of where in the different parts of the algorithmic value creation process bias might appear, see Silva & Kenney, 2019.

[x] Certainly, the impact of the array of digital tools on our social and economic lives is hard to capture, but one instance from our pandemic world of teleconnection struck us. Writing this essay in the Zoomed world of the pandemic, Mark reflected on the surveillance implications of Zoom while he was looking at a meeting screen with 230 people's video available, and the [REC] record light on in the upper-right. That is, the "host" was essentially capturing video surveillance information from 230 locations and recording it, and perhaps more relevant, the company Zoom could record and store on its servers all this surveillance data, and all with our implicit permission. "Cameras-on all-hands zoom meetings" has become a thing. Companies boast that they hold these meetings to show how they truly engage their staff. But if you bring it into the post-pandemic world, it would be the equivalent of calling an all-hands meeting in the auditorium, and kitting out the auditorium with, say, 1,500 cameras, one pointed at EACH SEAT. The possibilities of monitoring the company's workforce at work are evident, but the uses for the data extend beyond that. For example, if Amazon or a competitor like Wayfair, bought Zoom, the company, or contracted with Zoom, just to be able to get a sense for the decor in customer's various living spaces.

[xi] The OECD observatory has been promoted by Germany and the United States, and launched in February 2020. The United Nations panel was proposed by the Canadians and the French. They involve different sets of countries. It remains to be seen if they have different foci.

[xii] Regulators with general skills as much as specific domain knowledge will be required to evaluate particular sectoral approaches. Focusing on specifics requires that those in each application domain, or evaluating each application, have significant understanding of the AI tools if they are to evaluate the particular significance of the tools in the domain context and identify the consequences. That capacity, beyond the hand waving, may be, at least for the moment, difficult to replicate in multiple domains.

[xiii] There is a long debate, and literature, about how to promote innovation, about the links between innovation and economic growth. For our purposes and discussion, we simply set that aside.

[xiv] There are domain specific security concerns. Certainly, distinct security concerns emerge with retail or medical applications and in business-to-business applications. In retail or medical application, I may have concerns of what will happen to my data and who will use it for what purposes. It is a surveillance issue, if you will. In business-to-business applications, mechanisms for controlled sharing are often needed. If my equipment is part of a system application, I may want to protect my competitively sensitive data and intellectual property while allowing data required for the system to operate to be shared. Evidently, there are also systemic deployment concerns. We must be concerned that the hacking of basic infrastructure, cyber-attacks on the physical world, are all the more powerful when AI tools are deployed.

[xv] In the evaluation of AI, they are repurposing an argument and research work from the OECD (Andrews, Criscuolo, and Gal, 2016).

xvi. One might more bluntly ask how the losers, those disrupted or displaced, may be compensated, ignored, or suppressed. For an argument about this in earlier economic restructurings see Zysman, 1983.

[xvii] BRIE's work on digital platforms addresses these issues. Please view Berkeley Roundtable on the International Economy, 2020.

[xviii] The shift to an economy of services provided through digital portals, of digital services with everything, is explored in a set of BRIE publications: Zysman et al., 2013 and Nielsen, Murray, and Zysman, 2013.

[xix] Zysman and colleagues have argued this out thoroughly elsewhere. For example, see: Zysman, Kenney, and Tyson, 2019.

[xx] For example, see: Smit et al., 2020; Manyika et al., 2017; OECD, 2019; Nedelkoska and Quintini, 2018; and Arntz, Gregory, and Zierahn, 2016.

[xxi] Contrast Martin Ford and Ken Goldberg.

[xxii] Mark considers this issue by looking at his cappuccino maker. Mark notes that just because a machine or algorithm can be applied to perform a task, that does not guarantee that the machine is suitable to replace a human performing all aspects of the task. Yes, the cappuccino maker in the lab pretty much does the hard work (grinding beans, foaming milk, and driving pressured steam through the coffee mulch). But the machine, or its designer, outsources the task areas that are not cost effective for robots: filling the milk chiller from a carton in the nearby fridge, emptying the grounds, rinsing the drain plate, decalcifying the machine. Those are left to humans, the ultimate general purpose technologies.

[xxiii] Why do we argue that an upskilling deployment strategy, aimed at emphasizing business strategies and work organization tactics that support complementarity between technology and people, can encourage creating better jobs as well as new jobs?

Technology, the debate must recognize, is quite plastic. Scientific and technical advance creates a menu of possibilities. That the purposes and contexts of those deploying and implementing the particular applications shape the consequences of and often very character of technology is a view now supported by and entrenched in academic debate.

That plasticity is evident in both the technology and its deployment.

- Consider that the very character of software makes the deployment of technology and the skills required dependent on the interface between users and the underlying technology.
- User interfaces will affect the skills required to operate construction or mining equipment. Two contemporary examples: 1) intuitive user interfaces allow surgeons to use robots, or simple word processing programs augment the abilities of professors; 2) the use of a smartphone or a tablet or a notebook computer, all powerful digital systems, are facilitated by apps, user interfaces.
- The character of organization and organizational strategy produces very different approaches to the technology. Factories in firms adopting classic Fordist, rigid hierarchical arrangements differ in their uses of technology from firms with more flexible pragmatic managements. Retail stores in Denmark that assign significant responsibility to the shop floor differ sharply from Walmart-like arrangements.
    - In turn, the organizational differences often turn on attitudes to the value of the workforce, what workers know that is crucial and indispensable.

Work we underscore, is being transformed in the digital era. But the debate is wide open about how fast, when, how much will be displacement and how extensive will be reorganization, or why we already observe distinct national variation in both use of the technology and its impact on job distribution and equality. That ambiguity suggests there is room for choice. The Northern European countries, which depend heavily on skilled labor and those countries, such as Japan, that are facing skilled labor shortages are hypothesized to provide the richest search domain for positive examples.

[xxiv] "There is no U.S.-based wireless access equipment provider today that builds those solutions," said Sandra Rivera, a senior vice president at Intel who helps guide the chipmaker's 5G strategy (Fung, 2019).

REFERENCES

Andrews, D., Criscuolo, C., and Gal, P. (2016) 'The best versus the rest: The global productivity slowdown, divergence across firms and the role of public policy', *OECD Productivity Working Papers,* no. 5, https://doi.org/10.1787/63629cc9-en.

Arntz, M., Gregory, T., and Zierahn, U. (2016) 'The risk of automation for jobs in OECD countries: A comparative analysis', *OECD Social, Employment and Migration Working Papers*, No. 189, Paris: OECD Publishing.

Bank for International Settlements (2020) 'III. Central banks and payments in the digital era' *Annual Economic Report 2020*, pp. 67-95.

Bearson, D., Kenney, M., and Zysman, J. (2019) 'Labor in the platform economy: New work created, old work reorganized and value reconfigured', *BRIE Working Paper Series*.

Berkeley Roundtable on the International Economy (2020) 'The platform economy', University of California, Berkeley.

Blakeslee, S. (2006) 'Cells that read minds', *The New York Times,* 10 January.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., and Amodei, D. (2018) 'The malicious use of artificial intelligence: Forecasting, prevention, and mitigation', *arXiv*.

Christensen, C. (1997) *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail,* Brighton: Harvard Business Review Press.

Cutolo, D. and Kenney, M. (2020) 'Platform-dependent entrepreneurs: Power asymmetries, risks, and strategies in the platform economy', *Academy of Management Perspectives*.

Deutsch, M. (2016) 'Harry Potter: written by artificial intelligence', *Deep Writing, 8* July 8.

Eckersley, P. and Nasser, Y. (2017) 'AI progress measurement', *Electronic Frontier Foundation AI Progress Measurement Project*.

Fjelland, R. (2020) 'Why general AI will not be realized', *Humanities and Social Science Communications* 7(10): 1-9.

Friedersdorf, C. (2018) 'YouTube extremism and the long tail', *The Atlantic,* 12 March.

Fung, B. (2019) 'How China's Huawei took the lead over US companies in 5G technology', *Washington Post,* 10 April.

Funk, J. (2020) 'Expect evolution, not revolution: Despite the hype, artificial intelligence will take years to significantly boost economic productivity', *IEEE Spectrum* 57(3):30-35.

Google (2018) 'Artificial intelligence at Google: Our principles', *Google AI*.

Gopnik, G. (2011) 'What do babies think?', Presentation at TEDGlobal 2011.

Griffith, M., Steinberg, R., and Zysman, J. (2017) 'From great power politics to a strategic vacuum: Origins and consequences of the TPP and TTIP', *Cambridge University Press* 19(4): 573-592.

(2016) "Harold Cohn, artists – obituary", *Telegraph*, 22 May.

Hill, K. (2020) 'The secretive company that might end privacy as we know it', *The New York Times* 10 February.

Joshi, M., Choi, E., Weld, D.S., and Zettlemoyer, L. (2017) 'TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension', *Association for Computational Linguistics (ACL).*

Kenney, M., Bearson, D., and Zysman, J. (2019) 'The platform economy matures: Pervasive power, private regulation, and dependent entrepreneurs', *BRIE Working Paper Series*.

Kenney, M. (2020) 'The platform economy matures: Pervasive power, private regulation, and dependent entrepreneurs', Online presentation at the SASE Annual Conference, 21 July.

Kenney, M. and Zysman, J. (2016) 'The rise of the platform economy', *Issues in Science and Technology* 32(3): 61-69.

Knight, W. (2020) 'If AI's so smart, why can't it grasp cause and effect?', *WIRED,* 9 March.

Koebler, J. (2015) 'Elon Musk on superintelligent robots: We'll be lucky if they enslave us as pets', *Vice,* 23 March.

Lee, K. (2018) *AI Superpowers: China, Silicon Valley, and the New World Order,* Boston: Houghton Mifflin Harcourt.

Leslie, D. (2019) 'Raging robots and hapless human: The AI dystopia', *Nature* **574**: 32-33.

Lessig, L. (1999) *Code and other Laws of Cyberspace,* New York: Basic Books.

Lessig, L. (2000) 'Code is law: On liberty and cyberspace', *Harvard Magazine,* 1 January.

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., and Sanghvi, S. (2017) 'Jobs lost, jobs gained: Workforce transitions in a time of automation', *McKinsey Global Institute*.

Marcus, G. and Davis, E. (2019) *Rebooting AI: Building Artificial Intelligence We Can Trust.* New York: Pantheon Books, 7.

Markus, G. (2018) 'Deep learning: A critical appraisal', *arXiv*.

Marr, B. (2019) '13 mind-blowing things artificial intelligence can already do today', *Forbes,* 11 November.

McGee, P. and Chazan, G. (2020) 'The Apple effect', *The Financial Times*, 30 January.

Mitchell, M. (2019) *Artificial Intelligence: A Guide for Thinking Humans*, New York: Farrar Straus and Giroux.

Nedelkoska, L. and Quintini G. (2018) 'Automation, skills use and training' *OECD Social, Employment and Migration Working Papers* 202.

Nielsen, N. C., Murray, J., and Zysman, J. (2013) *The Services Dilemma: Productivity Sinkhole or Commoditization?* Harpenden: Verve Books.

OECD (2019) *OECD Employment Outlook 2019: The Future of Work*, Paris: OECD Publishing.

O'Neil, C. (2016) *Weapons of Math Destruction,* New York: Crown Books.

Russel, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*, New York: Viking Press.

Samuel, S. (2020) 'Kids' brains may hold the secret to building better AI', *Vox,* 28 February 28.

Silva, S. and Kenney, M. (2019) 'Algorithms, platforms, and ethnic bias', *Communications of the Association of Computing Machinery* 62(11): 37-39.

Smit, S., Tacke, T., Lund, S., Manyika, J., and Thiel, L. (2020) 'The future of work in Europe: Automation, workforce transitions, and the shifting geography of employment', *McKinsey Global Institute*.

Turca, D. (2016) ‘Is technology downskilling or upskilling us?’ *Boostzone Institute,* 12 June 12.

Tyson, L. (2019) ‘How automation will affect the future of work in Germany’, *The Berlin Journal*.

Zysman, J. (1983) *Governments Markets and Growth: Financial Systems and Politics of Industrial Change,* Ithica: Cornell University Press.

Zysman, J., Feldman, S., Kushida, K., Murray, J., Nielsen, N.C. (2013) ‘Services with Everything: The ICT-Enabled Digital Transformation of Services’, In D. Breznitz and J. Zysman (Eds.), *The Third Globalization: Can Wealth Nations Stay Rich in the Twenty-First Century?,* Oxford: Oxford Scholarship Online, pp. 99-129.

Zysman, J., Kenney, M., and Tyson, L. (2019) ‘Beyond hype and despair: Developing healthy communities in an era of intelligent tools’, Innovation Policy White Paper Series 2019-01, Munk School of Global Affairs, University of Toronto.