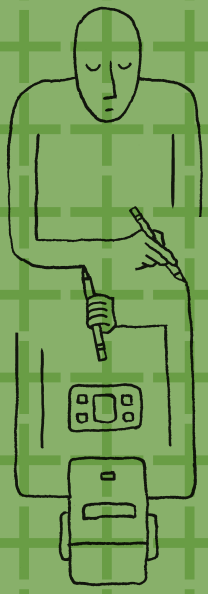
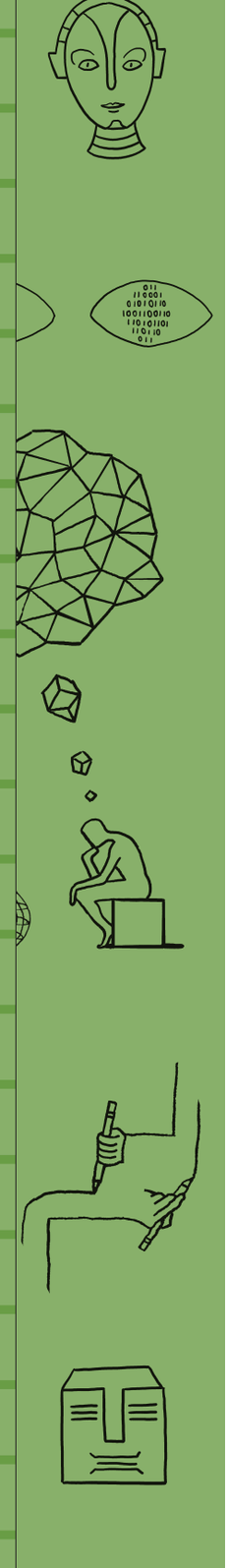
AI+1

Shaping our integrated future

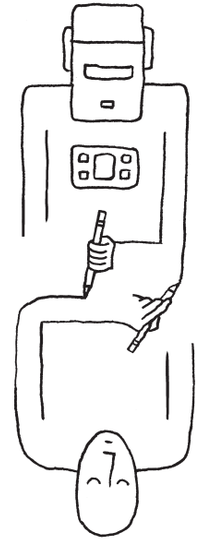


Join the conversation: **rockefellerfoundation.org**



AI+1

Shaping our integrated future



fore+word

I am delighted to invite you on an intellectual journey: a collection of thoughts, ideas and calls for action on artificial intelligence (AI), written by some of the world's most-promising minds in the field. In 2020, AI is drawing more and more attention and driving more and more conversations—and rightly so. It is high on the agendas of research institutions, nongovernmental organizations, business leaders and governments across the world. The focus on AI will only grow because the advance of the technology is unstoppable. But as a society, we have a responsibility to pause and think about its implications.

For The Rockefeller Foundation, grappling with the opportunities and challenges of this technological breakthrough is already a long-standing tradition. In 1956, we funded the Dartmouth Summer Research Project, an event that included renowned scientists, engineers, and mathematicians who coined the term *artificial intelligence* during that gathering. Its purpose was stated as follows:

"The study is to proceed on the basis of the conjecture that every aspect of learning or any feature of intelligence can, in principle, be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."

Decades later, AI now has greater relevance than perhaps ever before, and as a foundation dedicated to harnessing the frontiers of data, science, and technology for the good of all humanity, it is only natural that we continue to drive the discourse on AI in 2020. Yet today, our duty is not only to encourage technological breakthroughs; we must also ensure that the future of AI is designed *responsibly*.

Now, in the face of the immediate threat from COVID-19, and the longer-term, more intractable—but no less urgent—threats of poverty and climate change, we believe that the playbook that has served us for over a century, through pandemics, wars, and depressions, can serve us today.

Our charge is to ensure AI solves problems instead of creating new ones: to steer its progress toward new missions to alleviate chronic social challenges such as ill health, hunger, poverty, and inequality – instead of deepening them. This notebook of ideas before you is one more step in this direction. **Onward.**

Dr. Rajiv J. Shah – President, The Rockefeller Foundation
@rajshah

contents

an open invitation

to shape our integrated future

In October 2019, The Rockefeller Foundation hosted an exceptional group of technologists, philosophers, economists, lawyers, artists and philanthropists to explore how we can harness artificial intelligence (AI) to create a better future for humanity.

We met at the foundation's inspiring Bellagio Center on Lake Como in Italy to break away from our regular work and routines so we could slow down, connect, debate and create. It was not an easy conversation! But participants left with a wider aperture of the issues, a deeper understanding of others' perspectives and new relationships to draw upon in the future. If action is a team sport, participants drafted new players to join projects they were pursuing, including new ones that were born in just a few days—initiatives we are eager to see bear fruit in 2020 and beyond.

The convening was far from perfect. We didn't have every voice we needed in the room or enough time to explore every topic. Still, the discussions inspired new ideas and motivated collective actions. To share insights with the broader Rockefeller Foundation community and beyond, we wanted to impart messages that resonated at our meeting.

To convey the breadth of the conversation, we asked 14 participants to contribute essays to this report. Our goal was to surface common themes that are informing the foundation's future approach while maintaining the texture that gave the convening its richness. We're grateful for and excited by their contributions.

Our first essay, by Tim O'Reilly, founder and CEO of O'Reilly Media, Inc., sets the stage for the breadth of systems and issues in play if we want AI to become a force for good. O'Reilly indicates the need for new thinking around AI governance. Andrew Zolli, who oversees global impact initiatives at Planet, looks at these same system dynamics and governance questions through different lenses that examine the optimistic opportunity for AI to be geared towards human well-being and self-expression. Zolli focuses on how AI can integrate with human thinking to enhance our capabilities.

Marietje Schaake, international policy director at Stanford University's Cyber Policy Center, explores government's role in managing AI. A conversation between leading data scientists Jake Porway, founder and executive director of DataKind, and Hilary Mason, founder of Fast Forward Labs, digs into the interests of and opportunities with the private sector.

The diverse team of Katarzyna Szymielewicz, Richard Whitt—both of them lawyers, Whitt previously at Google—and technologist/designer Amir Baradaran explore AI governance and management from a human user's point of view, offering new models for third parties. And the CEO of the Cloudera Foundation, Claudia Juech, extends this user's perspective by evaluating the factors that ensure success for data science and AI projects in nonprofit settings.

Open Data Services Co-operative co-founder Tim Davies focuses on root cause issues related to the data that fuels AI systems. Stefaan Verhulst, cofounder of New York University's GovLab, posits that sharing private-sector data through new partnership models and roles makes possible the use of such data for public-good purposes—both the everyday and the urgent.

Turning to the need to expand the cognitive tools we use to address these questions, Nils Gilman, vice president at the Berggruen Institute, and Maya Indira Ganesh of Leuphana University's Digital and Media Studies department examine the assumptions underlying the metaphors we use to think about, talk about—and act on—AI, particularly when it comes to policies. We conclude with Maya's provocative piece which highlights the complex interplay between people and machines that we can expect as AI becomes more integrated into our daily lives.

Finally, in a visual piece, Sarah Newman interprets recent trends in AI to stimulate new and more open-ended ways of thinking about the issues.

Some pieces cover similar topics but from different angles. This surfaces the nuances involved. As we collaborate across disciplines and sectors to tackle the questions raised by AI's proliferation, we have to embrace nuance to avoid talking past each other in order to craft holistic solutions.

AI will eventually be ubiquitous in the background of day-to-day life, just like electricity. We need to shape AI as a technology that will weave together our integrated human + digital future.

we need to develop new governance approaches to ensure a responsible AI future

Recovering from the COVID-19 pandemic will soon create a new normal from which this integrated future will emerge. New models of AI governance need to be developed because current rules and rule-making systems are not up to the task. Much as the world developed the Bretton Woods system for managing newly complex global monetary relations after World War II, we need to develop new governance approaches to ensure a responsible AI future.

In the conclusion, Hunter Goldman, Director of innovation at The Rockefeller Foundation, offers initial thoughts on this direction.

We decided to call this *a notebook of ideas* because it is neither comprehensive nor prescriptive. It aims to raise a few key questions rather than give absolute answers.

Our writers explored the issues we know about today and the unknowns that will emerge tomorrow. We hope insights in this notebook spark novel ideas to ensure that AI serves the well-being of humanity throughout the world.

We look forward to you joining us on this journey.

Zia Khan – Senior Vice President of Innovation,
The Rockerfeller Foundation
@ziakhannyc

we have already let the genie out of the bottle

How will we make sure that Artificial Intelligence won't run amok and will be a force for good?

There are many areas where governance frameworks and international agreements about the use of artificial intelligence (AI) are needed.

For example, there is an urgent need for internationally shared rules governing autonomous weapons and the use of facial recognition to target minorities and suppress dissent. Eliminating bias in algorithms for criminal sentencing, credit allocation, social media curation and many other areas should be an essential focus for both research and the spread of best practices.

Unfortunately, when it comes to the broader issue of whether we will rule our artificial creations or whether they will rule us, we have already let the genie out of the bottle.

In his book *Superintelligence: Paths, Dangers, Strategies*, Nick Bostrom posited that the future development of AI could be a source of existential risk to humanity via a simple thought experiment. A self-improving AI, able to learn from its experience and automatically improve its results, has been given the task of running a paper clip factory. Its job is to make as many paper clips as possible. As it becomes superintelligent, it decides that humans are obstacles to its singular goal and destroys us all. Elon Musk created a more poetic version of that narrative, in which it is a strawberry-picking robot that decides humanity is in the way of "strawberry fields forever."

It is not artificial intelligence we most have to fear but artificial single-mindedness.

it's not artificial intelligence we have most to fear but artificial single-mindedness

What we fail to understand is that we have already created such systems. They are not yet superintelligent nor fully independent of their human creators, but they are already going wrong in just the way that Bostrom and Musk foretold. And our attempts to govern them are largely proving ineffective. To explain why that is, it is important to understand how such systems work. Let me start with a simple example. When I was a child, I had a coin-sorting piggy bank. I loved pouring in a fistful of small change and watching the coins slide down clear tubes, then arrange themselves in columns by size, as if by magic. When I was slightly older, I realized that vending machines worked much the same way and that it was possible to fool a vending machine by putting in a foreign coin of the right size or even the slug of metal punched out from an electrical junction box. The machine didn't actually know anything about the value of money. It was just a mechanism constructed to let a disk of the right size and weight fall through a slot and trip a counter.

If you understand how that piggy bank or coin-operated vending machine works, you also understand quite a bit about systems such as Google search, social media newsfeed algorithms, email spam filtering, fraud detection, facial recognition and the latest advances in cybersecurity. Such systems are sorting machines. A mechanism is designed to recognize attributes of an input data set or stream and to sort it in some manner. (Coins come in different sizes and weights. Emails, tweets and news stories contain keywords and have sources, click frequencies and hundreds of other attributes. A photograph can be sorted into cat and not-cat, Tim O'Reilly and not-Tim O'Reilly.) People try to spoof these systems—just like I and my teenage peers did with vending machines—and the mechanism designers take more and more data attributes into account so as to eliminate errors.

A vending machine is fairly simple. Currency changes only rarely, and there are only so many ways to spoof it. But content is endlessly variable, and so it is a Sisyphean task to develop new mechanisms to take account of every new topic, every new content source and every emergent attack.

Enter machine learning. In a traditional approach to building an algorithmic system for recognizing and sorting data, the programmer identifies the attributes to be examined, the acceptable values and the action to be taken. (The combination of an attribute and its value is often called a feature of the data.) Using a machine-learning approach, a system is shown many, many examples of good and bad data in order to train a model of what good and bad looks like. The programmer may not always know entirely what features of the data the machine-learning model is relying on; the programmer knows only that it serves up results that appear to match or exceed human judgment against a test data set. Then the system is turned loose on real-world data. After the initial training, the system can be designed to continue to learn.

content is endlessly variable

If you've used the facial recognition features of Apple or Google's photo applications to find pictures containing you, your friends or your family, you've participated in a version of that training process. You label a few faces with names and then are given a set of photos the algorithmic system is fairly certain are of the same face and some photos with a lower confidence level, which it asks you to confirm or deny. The more you correct the application's guesses, the better it gets. I have helped my photo application get better at distinguishing between me and my brothers and even, from time to time, between me and my daughters, until now it is rarely wrong. It recognizes the same person from childhood through old age.

A human-machine hybrid

In practice, the vast algorithmic systems of Google, Facebook and other social media platforms contain a mix of sorting mechanisms designed explicitly by programmers and newer machine-learning models. Google search, for instance, takes hundreds of attributes into account, and only some of them are recognized by machine learning. These attributes are summed into a score that collectively determines the order of results. Google search is now also personalized, with results based not just on what the system expects all users to prefer but also on the preferences and interests of the specific user asking a question. Social media algorithms are even more complex, because there is no single right answer. "Right" depends on the interests of each end-user and, unlike with search, those interests are not stated explicitly but must be inferred by studying past history, the interests of an individual's friends and so forth. They are examples of what financier George Soros has called *reflexive systems*, wherein some results are neither objectively true or false, but the sum of what all the system's users ("the market") believe.

Note that these systems are hybrids of human and machine—not truly autonomous. Humans construct the mechanism and create the training data set, and the software algorithms and machine-learning models are able to do the sorting at previously unthinkable speed and scale. And once they have been put into harness, the data-driven algorithms and models continue not only to take direction from new instructions given by the mechanism designers but also to learn from the actions of their users.

> The individual machine components cannot be thought of as intelligent, but these systems as a whole are able to learn from and respond to their environment, to take many factors into account in making decisions and to constantly improve their results based on new information.

That's a pretty good definition of intelligence, even though it lacks other elements of human cognition such as self-awareness and volition. Just as with humans, the data used in training the model can introduce bias into the results. Nonetheless, these systems have delivered remarkable results—far exceeding human abilities in field after field.

In those hybrid systems, humans are still nominally in charge, but recognition of and response to new information often happens automatically. Old, hand-coded algorithms designed by human programmers are being replaced by machine-learning models that are able to respond to changes in vast amounts of data long before a human programmer might notice the difference. But sometimes the changes in the data are so significant—for example, makeup designed specifically to fool facial recognition systems, astroturfed content produced at scale by bots masquerading as humans or deepfake videos—that humans need to build and train new digital subsystems to recognize them. In addition, the human mechanism designers are always looking for ways to improve their creations.

Govern not by rules but by outcomes

The decades of successful updates to Google search in order to maintain search quality in the face of massive amounts of new information, adversarial attacks and changes in user behavior—as well as other success stories like antispam and credit-card-fraud-detection systems—provide some basis for understanding how to govern the AI of the future. Human control is expressed not through a fixed set of rules but through a set of desired outcomes. The rules are constantly updated in order to achieve those outcomes. Systems managed in this way represent a sharp break with previous, rules-based systems of governance.

Any governance system that tries to define, once and for all, a set of fixed rules is bound to fail. The key to governance is the choice of desired outcome, measurement of whether or not that outcome is being achieved and a constantly updated set of mechanisms for achieving it.

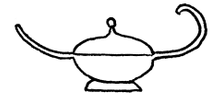
There are two levels of AI governance:

1. The microgovernance of constant updates in response to new information, expressed by building better algorithms and models
2. The macro-governance of the choice of outcome for which algorithms and models are optimized

Today's technology companies have gotten pretty good at level 1. Where they struggle is at level 2. The outcomes-based approach to governance does have an Achilles' heel. Algorithmic systems are single-minded optimizers. Much like the genies of Arabian mythology, they do exactly what their masters ask of them regardless of the consequences, often leading to unanticipated and undesirable results.

Peter Norvig, Google's director of research and co-author of the leading textbook on AI, notes that part of the problem is that it is hard to say what you want in a succinct statement—whether that statement is made in everyday language, legalese or a programming language. This is one advantage of machine learning over traditional systems. We show these systems examples of what we consider good and bad rather than try to summarize them once and for all in a single statement—much as human courts rely on case law.

> any governance system that tries to define, once and for all, a set of fixed rules is bound to fail



Another part of the problem is the hubris of thinking it is possible to give the genie a coherent wish. Norvig points out that we should recognize there will be errors and that we should use principles of safety engineering. As he said to me, "King Midas would have been OK if only he had said, 'I want everything I touch to turn to gold, but I want an undo button and a pause button.'"

I'm not sure that would be sufficient, but it's a good start.

Be careful what you ask for

If the outcome is well chosen and directed by the interests not only of the mechanism designer and owner but also of the system's users and society as a whole, the benefits can be enormous. For example, Google set as its corporate mission "to organize the world's information and make it universally accessible and useful." Few could deny that Google has made enormous progress toward that goal. But in a hybrid system, goals set in human terms must be translated into the mathematical language of machines. That is done through something referred to as an *objective function*, whose value is to be optimized (maximized or minimized.) Google's search algorithms are relentlessly optimized for producing answers—originally, a list of pointers to websites and now, for many searches, an actual answer—that satisfy users, as measured by the fact that they go away and don't make the same search a second time.

Facebook, too, lays claim to a noble mission. It aims "to give people the power to build community and bring the world closer together." However, in fulfillment of that mission, Facebook tasked its systems with optimizing for what might broadly be called *engagement*, measuring such factors as how much time users spend on the site and how many posts they read, like and respond to. The system's creators believed that this would bring their users closer relationships with their friends, but we now know that instead, it drove divisiveness, addictive behavior and a host of other ills. Not only that, but outsiders learned to game the system in order to manipulate Facebook's users for their own ends.

Like other social media platforms, Facebook has made progress at the microlevel of governance by targeting hate speech, fake news and other defects in the newsfeed curation performed by its algorithmic systems in much the same way that the vending machines of my childhood added new tests to avoid dispensing candy bars in exchange for worthless metal slugs.

if the outcome is well chosen and directed by the interests not only of the mechanism designer and owner but also of the system's users and society as a whole the benefits can be enormous

But the company is still struggling with the higher-level question of how to express the human wish to bring people together in a mathematical form that will cause its genies to produce the desired outcome.

Most troubling is the question, What are Facebook's alternatives if greater engagement with its services is not actually good for Facebook users? The value of the company depends on growth in users and usage. Its advertising premium is based on microtargeting, wherein data about users' interests and activities can be used for their benefit but can also be used against them. In far too many cases, when the interests of its users and the interests of its advertisers diverge, Facebook seems to take the side of the advertisers—even to the point of knowingly accepting false advertising over the protests of its own employees.

The problem of mixed motives

Despite its history of success in constantly updating its search engine for the benefit of its users, Google fell prey to many of the same problems as Facebook at its YouTube unit. Unable to use "give them the right answer and send them away" for the majority of its searches, YouTube chose instead to optimize for time spent on the site and ended up with disinformation problems at least as bad as those that bedevil Facebook—and quite possibly worse.

Even at its search engine unit, Google seems to have turned away from the clarity of its original stance on the divergence of interest between its users and its advertisers, which Google cofounders Larry Page and Sergey Brin had identified in their original 1998 research paper on the Google search engine. In an appendix titled "Advertising and Mixed Motives," they wrote, "We expect that advertising-funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers."

At the time, Page and Brin were arguing for the existence of an academic search engine without commercial motives as a check on that problem. But with the adoption of pay-per-click advertising, whereby advertisers are charged only when a user clicks on an ad—presumably because the user found it useful—they believed they had found a way to align the interests of the company's two prime constituencies. In the company's first decade or so, Google also made a clear separation between the systems that served its end-users and the systems that served its advertisers. Ad results were calculated separately and shown completely separately from organic search results. But gradually, the boundaries began to blur. Ads, formerly in a secondary position on the page and highlighted in a different color, began to take on more and more prominent positions and to become less distinguishable from organic search results.

Google also seems to have re-evaluated the relationship between itself and the content suppliers of the World Wide Web. The company began as a way to match information seekers with information providers in this vast new marketplace for human collective intelligence. Its job was to be a neutral

middleman, using its massive technological capabilities to search through what was to become trillions of Web pages in order to find the page with the best answer to trillions of searches a year. Google's success was measured not only by the success of its users but also by the success of the other sites that the search engine sent the users off to.

In an interview that was attached to the Form S-1 filing for Google's 2004 IPO, Page said: "We want you to come to Google and quickly find what you want. Then we're happy to send you to the other sites. In fact, that's the point. The portal strategy tries to own all of the information… Most portals show their own content above content elsewhere on the Web. We feel that's a conflict of interest—analogous to taking money for search results. Their search engine doesn't necessarily provide the best results; it provides the *portal's* results. Google conscientiously tries to stay away from that. We want to get you out of Google and to the right place as fast as possible. It's a very different model."

Page and Brin seem to have understood at the time that success did not mean success for only themselves—or even for their customers and advertisers—but for the ecosystem of information providers whose content Google had been created to search. Google's early genius was in balancing the competing interests of all those different constituencies. This is the positive future of AI. As Paul Cohen, former DARPA program manager of AI who is now dean of the School of Computing and Information at the University of Pittsburgh, once said, "The opportunity of AI is to help humans model and manage complex, interacting systems," yet 15 years after Page said Google's aim was to send users on their way, more than 50% of all searches on Google end on Google's own information services, with no click-through to third-party sites; and for any search that supports advertising, paid advertising has driven organic search results far below the fold.

What is going on here?

There are two answers, and both of them shed light on the governance problem for AI. The first is that the humans in charge of directing the AI may change their idea about what they want—even while telling themselves that their wishes have not changed. Google's mechanism designers tell themselves it is better for Google users to simply get answers than be connected to an external Web page. And they may well be right about that. But in the way they chose to implement this, they are no longer indisputably honest brokers. Their own content pages appear at the top of many searches—exempt from the algorithmic systems that use data to find which pages their users actually consider to be the best. The ads Google runs are now seen before rather than beside organic search results.

In 1930, Upton Sinclair gave a cogent explanation for what appears to be a change in Google's human managers' understanding of their own goals: "It is difficult to get a man to understand something when his salary depends on his not understanding it!" That "mixed motive," as Page and Brin originally described it, becomes increasingly dangerous as algorithmically managed platforms become more dominant. They may use their power to favor themselves rather than their customers. It is also possible to build systems that actually intend harm—the way that repressive regimes around the world today track minorities and suppress dissent. As writer and activist Zeynep Tüfekçi tweeted, "Too many worry about what AI—as if some independent entity—will do to us. Too few people worry what 'power will do with' AI."

systems and architectures that distribute rather than centralize the power of AI may well be key to its governance

For that reason, systems and architectures that distribute rather than centralize the power of AI may well be key to its governance. As Adam Smith famously wrote, "It is not from the benevolence of the butcher, the brewer or the baker that we expect our dinner but from their regard to their own interest." So, too, argues AI pioneer Stuart Russell in his new book, *Human Compatible: Artificial Intelligence and the Problem of Control*, it is not from the benevolence of a single centralized AI that we should expect our prosperity but from building systems that enable each of us to more fully express and pursue our own interests and preferences. The question of governing AI, in this sense, is the question of how to best govern human society as a whole with the aid of these new tools.

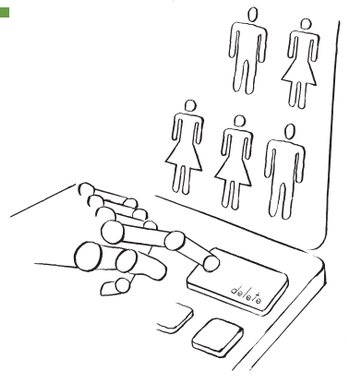
The second answer, which is even more alarming, is the prospect that even at these great and powerful companies, the humans are not really in charge. The human CEOs of companies are nominally governed by their human boards of directors, but in truth, they are governed by something called "the market"—a vast, reflexive algorithmic system in which companies are relentlessly tasked to optimize for growth and profits, even if human values must be ignored.

Our algorithmic master

It is not just companies like Google and Facebook that have moved from being traditional, human-directed organizations into a new kind of human–machine hybrid that ties their employees, their customers and their

suppliers into a digital, data-driven, algorithmic system. It is all companies whose stock is traded on public markets. Science fiction writer Charlie Stross calls modern corporations "slow AIs." And like Bostrom's paper clip maximizer, these AIs are already executing an instruction set that tells them to optimize for the wrong goal and to treat human values as obstacles.

How else can you explain a system that treats employees as a cost to be eliminated and customers and communities as resources to be exploited? How else can you explain pharmaceutical companies that worked consciously to deceive regulators about the addictiveness of the opioids they were selling, thereby triggering a devastating health crisis? How else can you explain decades of climate denial by fossil fuel companies, decades of cancer denial by tobacco companies and the obscene accounting manipulations and government derelictions to avoid paying the taxes that keep their host countries functioning? It is the machine that is in charge, and like all such machines, it thinks only in mathematics, with an objective function whose value is to be maximized.



Those who gave our companies and our markets the objective function of increasing shareholder value above all else believed that doing so would lead to greater human prosperity. When, in 1970, Milton Friedman wrote that the only social responsibility of a corporation is to increase its profits, he believed that that would allow shareholders, as recipients of those profits, to make their own determinations about how best to use them. He didn't imagine the race to the bottom of declining wages, environmental degradation and social blight that the single-minded pursuit of corporate profit would actually deliver. But after 1976, when Michael Jensen and William Meckling made the case that the best mechanism design for maximizing shareholder value was to pay executives in company stock, the human managers were made subservient to the objective of the machine.

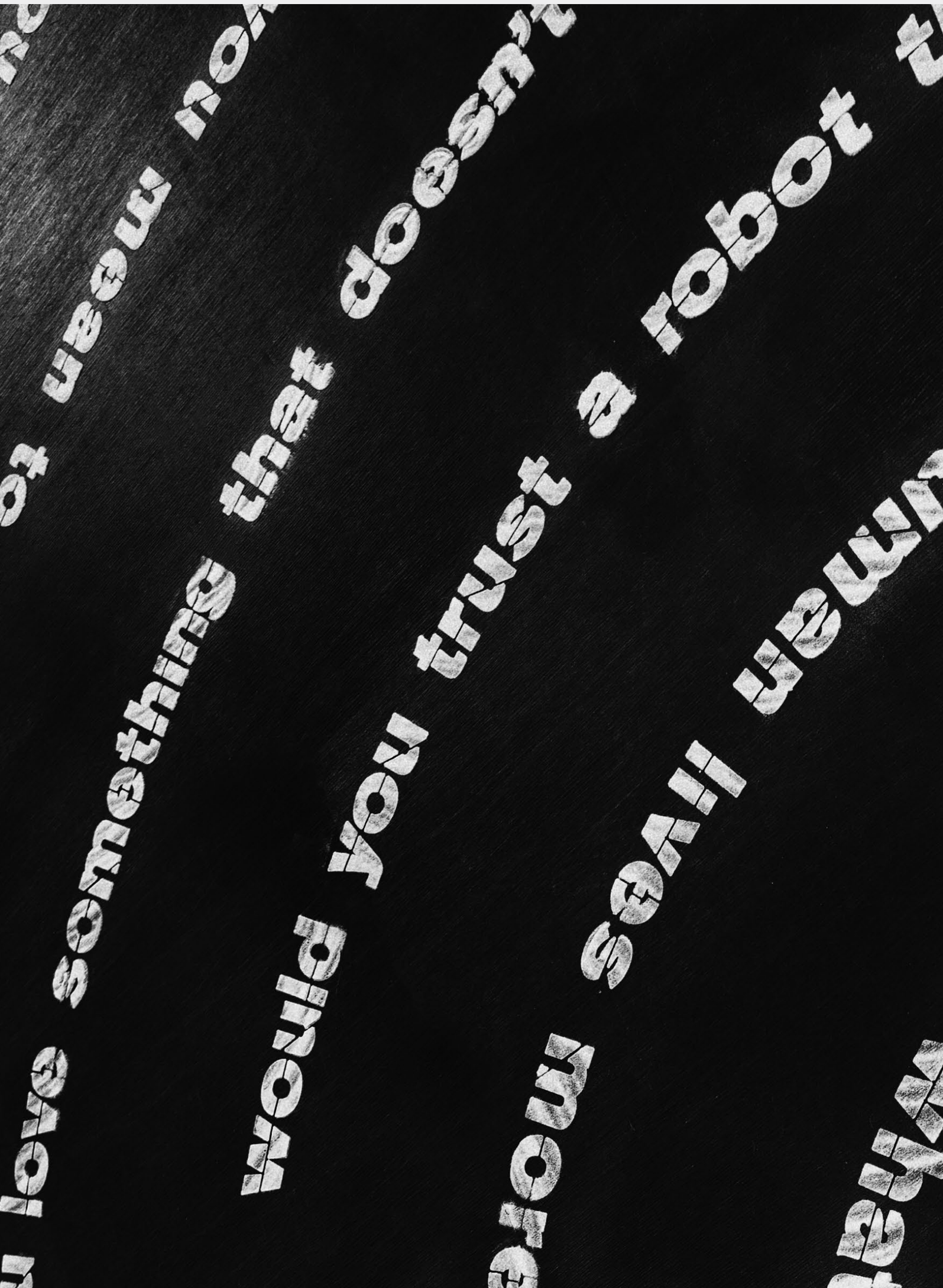
We now know that Friedman, Jensen and Meckling were wrong about the results they expected, but the mechanism has been built and enshrined into law. Those who designed it have passed on, and those who are now nominally in charge (our policy-makers, our economic planners, our legislators and our government executives) no longer entirely understand what was built or can no longer agree on how to change it. Government, too, has become a slow AI. As E. M. Forster wrote in *The Machine Stops*, "We created the Machine to do our will, but we cannot make it do our will now … We only exist as the blood corpuscles that course through its arteries, and if it could work without us, it would let us die." And so the paper clip maximizer continues its work, just as it has been commanded.

We humans do what we can to blunt this relentless command from our former algorithmic servant, now through ignorance and neglect, our algorithmic master. We adopt high-minded principles like those articulated by the Business Roundtable, promising to take into account not just corporate profit but also the needs of employees, customers, the environment and society as a whole. Attempts at governance of this kind are futile until we recognize that we have built a machine and set it on its course. Instead, we pretend that the market is a natural phenomenon best left alone, and we fail to hold its mechanism designers to account. We need to tear down and rebuild that machine, reprogramming it so that human flourishing, not corporate profit, becomes its goal. We need to understand that we can't just state our values. We must implement them in a way that our machines can understand and execute.

we need to tear down and rebuild that machine, reprogramming it so that human flourishing, not corporate profit, becomes its goal

And we must do so from a position of profound humility, acknowledging our ignorance and our likelihood to fail. We must build processes that not only constantly measure whether the mechanisms we have built are achieving their objective but that also constantly question whether that objective is the correct expression of what we actually desire. But even that may not be enough. As Russell notes in *Human Compatible*, the machinery we create must operate on the principle that it does not know the right objective. If it has a single-minded objective, a truly self-aware AI might well work to prevent us from changing it—and from detecting the need to change it—and so, our oversight is not enough.

The governance of AI is no simple task. It means rethinking deeply how we govern our companies, our markets and our society—not just managing a stand-alone new technology. It will be unbelievably hard—one of the greatest challenges of the twenty-first century—but it is also a tremendous opportunity.


would you trust a robot that doesn't
mean to love something that
would you trust a robot that
sarah+newman lives more

humanity and AI: conflict, co-operation, co-evolution

We shape our tools, and then they shape us.

"We shape our tools; thereafter, they shape us," noted media scholar John Culkin. It follows that evermore-powerful tools shape us in evermore-powerful ways—and few, if any, promise to do so as deeply as artificial intelligence.

There is no word in English for the dizzying mix of fascination, yearning and anxiety that mark contemporary discussions of AI. Every position is taken: Entrepreneurs wax breathlessly about its promise. Economists ponder its potential for economic dislocation. Policy-makers worry about reining in its potential for abuse. Circumspect engineers will tell you how much harder it is to implement in practice than headlines suggest. Activists point out AI's ability to quietly lock in our unconscious (and not so unconscious) biases, even while some of their peers are busy applying AI to try to overcome those very same biases. Techno-utopians look forward to an inevitable, ecstatic merger of humanity and machine. Their more cynical contemporaries worry about a loss of human control.

This diversity is fitting, for all of these positions are likely well-founded to some degree. AI *will* increase wealth, and concentrate wealth, and destroy wealth—all at the same time. It will amplify our biases and be used to overcome them. It will support both democracy and autocracy. It will be a means of liberation from, and subjugation to, various forms of labor. It will be used to help heal the planet and to intensify our consumption of its resources. It will enhance life and diminish it. It will be deployed as an instrument of peace and as an instrument of war. It will be used to tell you the truth and to lie to you. As folk singer Ani DiFranco observed, *every tool is a weapon, if you hold it right.*

> One thing we know: AI will not enter the scene neutrally. Its emergence will be conditioned as much by our cultural values, our economic systems, and our capacity for collective imagination as by the technology itself.

Is work drudgery or dignity? Who should decide what gets built? What should we *not* do, even if we can? Who matters? And what do we owe one another, anyway? How we answer these kinds of questions—indeed, whether we ask them at all—hints at the tasks we might assign to artificial intelligence and the considerations that will guide and constrain its emergence.

Between agony and opulence

Here in the West, AI is emerging at a moment of enormous imbalance of power between the techne—the realm of the builders and makers of technology—and the polis: the larger, social order. Techne has merged with the modern market and assumed, in places, both its agenda and its appetites.

An accelerating feedback loop is under way: powerful algorithms, deployed by evermore powerful enterprises, beget greater usage of certain digital products and platforms, which in turn generate ever-larger volumes of data, which inevitably are used to develop evermore effective algorithms; and the cycle repeats and intensifies. In the guise of useful servants, AI algorithms are pushing into every crevice of our lives, observing us, talking to us, listening to us. Always on, in the background, called forth like a djinn with a magic phrase: are they learning more about us than we know about ourselves? or misunderstanding us more than the deepest cynic might? Which is worse?

Those who control these algorithms and the vast troves of data that inform them are the new titans. Of the ten richest Americans, eight are technologists with a significant stake in the AI economy.[1] Together they own as much wealth as roughly half of humanity.

> “AI is something that will happen to my community,
>
> not for it

At the other end of the socioeconomic spectrum, in communities that receive these technologies, conversations about AI and automation are often colored by a pessimistic, rise-of-the-robots narrative that presupposes inevitable human defeat, downshifting and dislocation. “AI is something that will happen *to* my community,” a friend in an American Rust Belt city recently told me, “not *for* it.”

In this telling, the Uber driver’s side hustle—itself a response to the loss of a prior, stable job—is just a blip, a fleeting opportunity that will last only as long as it takes to get us to driverless cars, and then, good luck, friend! This is the inevitable, grinding endpoint of a worldview that frames technology primarily as a tool to maximize economic productivity, and human beings as a cost to be eliminated as quickly as possible. “Software is eating the world,” one well-known Silicon Valley venture capital firm likes to glibly cheer-lead, as if it were a primal force of nature and not a choice. How different the world looks to those whose livelihoods are on the menu.

Of course, it doesn’t have to be this way. Rather than deploying AI solely for efficient economic production, what if we decide to unleash its potential for the achievement of human well-being and self-expression? What if we used AI to narrow the gap between the agony and opulence that define contemporary capitalism? How might we return an AI dividend to citizens—in the form of reduced, more dignified, and more fulfilling labor and more free time?

Industrial revolutions are lumpy affairs; some places boom and others limp along. How might we smooth out the lumps? Maybe, as Bill Gates mused, if a robot takes your job, a robot should pay your taxes. (There’s a reason that Silicon Valley elites have recently become smitten with ideas of universal basic income: they know what’s coming.)

Ties that bind—and sever

But not just new economic thinking will be required. In a world where a small constellation of algorithmic arbiters frame what you see, where you go, whom you vote for, what you buy and how you are treated, threats to critical thinking, free will and social solidarity abound. We will use AI to shape our choices, to help us make them, and, just as often, to eliminate them. The more of our autonomy we cede to the machines, the more dependent we may become.

We are also just now beginning to understand how the algorithms that power social media amplify certain communities, discourses, politics and polities while invisibly suppressing others.

Liberal democracies are, at their core, complex power-sharing relationships, designed to balance the interests of individuals, communities, markets, governments, institutions and the rest of society's messy machinery. They require common frames of reference to function, rooted in a connective tissue of consensual reality. No one is really sure whether our algorithmically driven, hypertargeted social media bubbles are truly compatible with democracy as we have understood it. (That's a point well understood by both Cambridge Analytica, which weaponized AI to shatter the commons, and white nationalists, who've sought to legitimize and normalize their long-suppressed ideologies amid the shards. Both exploited precisely the same techniques.)

the more of our autonomy
we cede to the machines
the more dependent we may become

And all this is before so-called deepfakes—AI forgeries that synthesize apparent speech from well-known figures—and other digital chicanery are released at scale. There has been much hand-wringing already about the propaganda dangers of deepfakes, but the true power of such weaponized misinformation may, paradoxically, not be in making you believe an outright lie. Rather, it may simply suffice that a deepfake nudges you to feel a certain way—positively or negatively—about its subject, even when you know it's not real. Deepfakes intoxicate because they let us play out our pre-existing beliefs about their subjects as we watch. *What a buffoon!* or *That woman is a danger!* They trip our most ancient neural circuits—the ones that adjudicate in-groups and out-groups, us and them, revulsion and belonging. As such, AI may be used to harden

AI can and will also be used
to enrich the human spirit
expand our creativity and amplify
the true and the beautiful

lines of difference where they should be soft and to make the politics of refusal—of *de*consensus and dropping out—as intoxicating as the politics of consensus and coalition building.

Meanwhile, it is inevitable that the algorithms that underwrite what's left of our common public life will become increasingly politically contested. We will fight over AI. We will demand our inclusion in various algorithms. We will demand our exclusion from others. We will agitate for proper representation, for the right to be forgotten and for the right to be remembered. We will set up our own alternatives when we don't like the results. These fights are just beginning. Things may yet fall apart. The center may not hold.

Nudging our better angels

And yet, even though all of these concerns about politics and economics are legitimate, they do not tell anything like the complete story. AI can and will also be used to enrich the human spirit, expand our creativity and amplify the true and the beautiful. It will be used to encourage trust, empathy, compassion, co-operation and reconciliation—to create sociable media, not just social media.

Already, researchers have shown how they can use AI to reduce racist speech online, resolve conflicts, counter domestic violence, detect and counter depression and encourage greater compassion, among many other ailments of the human soul. Though still in their infancy, these tools will help us not only promote greater well-being but also demonstrate to the AIs that observe human nature just how elastic human nature is. Indeed, if we don't use AI to encourage the better angels of our nature, these algorithms may come to encode a dimmer view and, in a reinforcing feedback loop, embolden our demons by default.

An abundance of prediction, a scarcity of wisdom

AI will not only emulate human intelligence; it will also transform what it means for people to perceive, to predict and to decide.

When practical computers were first invented—in the days when single machines took up an entire floor of a building—computation was so expensive that human beings had to ration their own access to it. Many teams of people would share access to a single computational resource, even if that meant running your computer program at four in the morning.

> with enough sensors and enough data, the algorithms of AI will shift us from a real-time understanding to an increasingly predictive understanding of the world

Now, of course, we've made computation so absurdly cheap and abundant that things have reversed: it's many computers that share access to a single person. Now, we *luxuriate* in computation. We leave computers running, idly, doing nothing in particular. We build what are, in historical terms, frivolities like smart watches and video games and mobile phones with cameras optimized to let us take pictures of our breakfasts. We expand the range of problems we solve with computers and invent new problems to solve that we hadn't even considered problems before. In time, many of these frivolities have become even more important to us than the serious uses of computers that they have long-since replaced.

The arrival of AI is fostering something deeply similar—not in the realm of computation but in its successors: measurement and prediction.

As we instrument the world with more and more sensors, producing ever more data and analyzing them with evermore powerful algorithms, we are lowering the cost of measurement. Consequently, many more things can be measured than ever before. As more and more of the world becomes observable with these sensors, we will produce an ever-increasing supply of indicators, and we will move from a *retrospective* understanding of the world around us to an increasingly complete *real-time* one. Expectations are shifting accordingly. If the aperture of contemporary life feels like it's widening and the time signature of lived experience feels like it's accelerating, this is a reason.

And this is mere prelude. With enough sensors and enough data, the algorithms of AI will shift us from a real-time understanding to an increasingly predictive understanding of the world: seeing not just what was or what is but also what is likely to be.

Paradoxically, in many fields, this will likely *increase* the premium we put on human judgment—the ability to adeptly synthesize this new bounty of indicators and make sound decisions about them. An AI algorithm made by Google is now able to detect breast cancer as well as or better than a radiologist can;[2] and soon, others may be able to predict your risk of cancer many years from now. Still, it's your oncologist who is going to have to synthesize these and dozens of other signals to determine what to do in response. The more informed the doctor's decisions become, the more expensive they are likely to remain.

Eventually, machines will augment—or transcend—human capabilities in many fields. But that is not the end of the story. You can see that in the domains where AI has been deployed the longest and most impactfully. There *is* a story after the fall of man.

Consider what has happened in perhaps the *ur*-domain of artificial intelligence: chess.

When IBM's Deep Blue computer beat Garry Kasparov in 1997, ending the era of human dominance in chess, it was a John-Henry-versus-the-steam-engine-style affair. A typical grand master is thought to be able to look 20 or 30 moves ahead during a game; a player of Kasparov's exquisite skill might be expected to look substantially farther than that. Deep Blue, however, was able to calculate 50 *billion* possible positions in the three minutes allocated for a single move. The chess master was simply computationally outmatched.

After the fall

Deep Blue's computational advantage wasn't paired with any deep understanding of chess as a game, however. To the computer, chess was a very complex mathematical function to be solved by brute force, aided by thousands of rules that were artisanally hand-coded into the software by expert human players. Perhaps unsurprisingly, Deep Blue's style of play was deemed "robotic" and "unrelenting." And it remained the dominant style of computational chess in Deep Blue's descendants, all the way to present day.

All of that changed with the recent rise of genuine machine-learning techniques proffered by Google's DeepMind unit. The company's AlphaZero program was given only the rules of chess—and then played itself: 44 million times. After just four hours of self-training and playing itself,

it was able to develop sufficient mastery to become the most successful chess-playing entity—computer or human—in history.

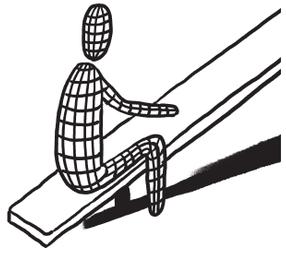
Several things are notable about AlphaZero's approach. First, rather than evaluating tens of millions of moves, the program analyzed only about 60,000—approaching the much more intuitive analysis of human beings rather than the brute-force methods of its predecessors.

Second, the *style* of AlphaZero's play stunned human players, who described it as "beautiful," "creative" and "intuitive"—words that one would normally associate with human play. Here was a machine with an apparently deep understanding of the game itself, evidencing something very close to human creativity. Being self-taught, AlphaZero was unconstrained by the long history of human styles of play. It discovered not only our human strategies—and by itself!—but also entirely new ones—ones never seen before.

> Here is the fascinating, *deeper* lesson: after a long age of human dominance in a particular intellectual pursuit falls before AI, *we don't turn away from those pursuits where we have been bested.*

It's possible that AlphaZero's successors may one day develop strategies that are fundamentally incomprehensible to us; but in the meantime, they are *magnificent* teachers that are expanding humanity's understanding of the truth of the game in a way no human grand master could. Even the programs that AlphaZero bested—those brute-force approaches that are themselves better than any human player's—are, through their wide availability, improving everyone's game. Somehow, our diminished status doesn't reduce our love of chess—much in the way that the reality of LeBron James doesn't diminish our love of playing basketball.

the age of human dominance in some fields will come to a close, as it already has in many areas of life



Variations of this story will unfold in every field and creative endeavor. Humanity will be stretched by artificial intelligence, augmented and empowered by it and, in places, bested by it. The age of human dominance in some fields will come to a close, as it already has in many areas of life. That will be cause for concern but also for celebration, because we humans admire excellence and we love to learn; and the rise of AI will provide ample opportunities for both.

Artificial intelligence will provide us with answers we couldn't have arrived at any other way. We can ensure a *humane* future with AI by doing what we do best: relentlessly asking questions, imagining alternatives and remembering the power inherent in our choices. We have more than we know.

that will be cause for concern but also for celebration

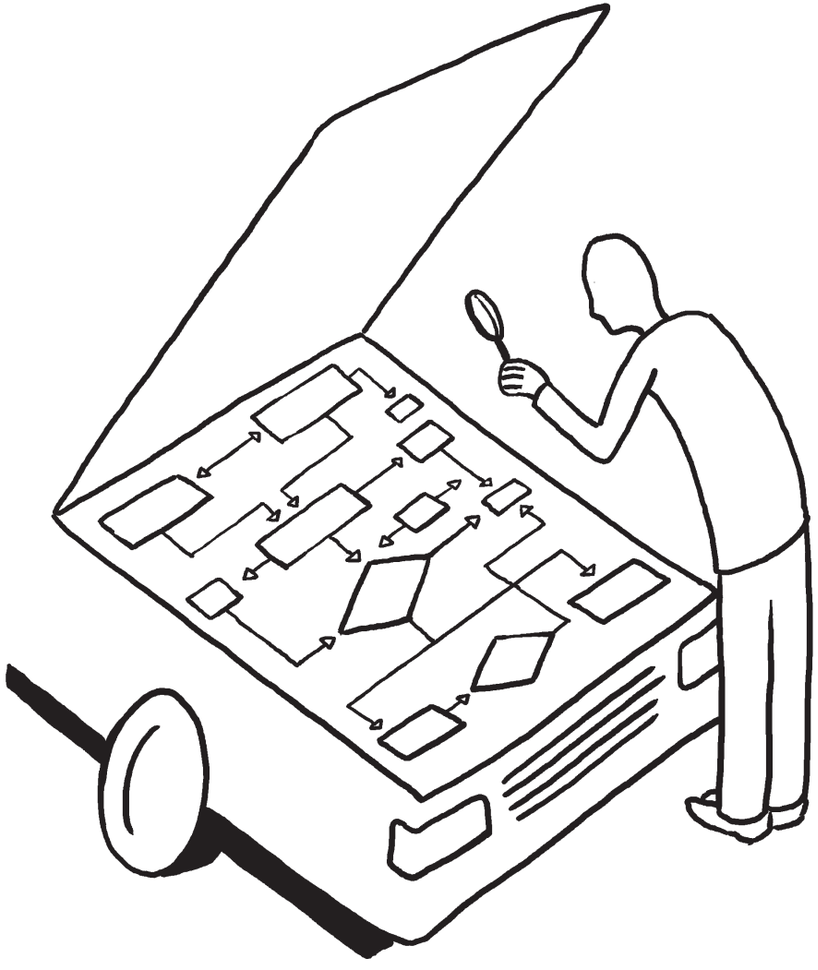
1. https://www.usatoday.com/story/money/2019/10/03/forbes-400-amazon-ceo-remains-richest-person-us-despite-divorce/3849962002/.
2. https://www.healthimaging.com/topics/artificial-intelligence/google-ai-rivals-radiologists-breast-cancer.

machine becomes sentient!
hurt a robot if it could feel pain
is a tee hee moral
sometimes

AI's invisible hand:

why democratic institutions need more access to information for accountability

marietje+schaake



Ethics and self-regulation are not enough

Across the world, artificial intelligence (AI) elicits both hope and fear. AI promises to help find missing children and cure cancer. But concerns over harmful AI-driven outcomes are equally significant. Lethal autonomous weapon systems raise serious questions about the application of armed-conflict rules. Meanwhile, anticipated job losses caused by automation top many governments' agendas. Effectively, AI models govern significant decisions impacting individuals, but they also ripple through society at large.

Yet discussions of AI's likely impact cannot be only binary—focused on gains and losses, costs and benefits. Getting beyond hope and fear will require a deeper understanding of AI-application-triggered decisions and actions amid their intended and unintended consequences. The troubling reality, however, is that the full impact of the massive use of tech platforms and AI is still largely unknown. But AI is too powerful to remain invisible.

the proper authorities must have the freedom to look under the algorithmic hood

Access to information forms the bedrock of many facets of democracies and the rule of law. Facts inform public debate and evidence-based policy making. Scrutiny by journalists and parliamentarians and oversight by regulators and judges require transparency. But private companies keep crucial information about the inner workings of AI systems under wraps. The resulting information gap paralyzes lawmakers and other watchdogs, including academics and citizens who are unable to know of or respond to any AI impacts or missteps. And even with equal access to proprietary information, companies examine data through different lenses and with different objectives than those used by democratic institutions which serve and are accountable to the public.

The starting-point for AI debates is equally flawed. Such conversations often focus on outcomes we can detect. Unintended consequences such as bias and discrimination inadvertently creep into AI algorithms, reflecting our offline world, or erroneous data sets and coding. Many organizations focus on correcting the damage caused by discriminatory algorithms. Yet we must know what we may expect from AI when it works exactly as anticipated. Before addressing the sometimes discriminatory nature of facial recognition technologies, we need to know if the technologies respect the right to privacy.

But AI and new technologies disrupt not only industries. They also systemically disrupt democratic actors' and institutions' ability to play their respective roles.

We must devote more attention to actors' and institutions' ability to access AI. This is a precondition for evidence-based regulation.

The key to the algorithmic hood

AI engineers admit that no one knows where the heads and tails of algorithms end after endless iterations. But we can know AI's unintended outcomes only when we know what was intended in the first place. This requires transparency of training data, documentation of intended outcomes and various iterations of algorithms. Moreover, independent regulators, auditors and other public officials need mandates and technical training for meaningful access to, and understanding of, algorithms and their implications.

Accountability is particularly urgent when AI-based, government-provided systems are used for tasks or services that encroach into the public sphere. Such outsourced activities include the building and defense of critical infrastructure, the development and deployment of taxpayer databases, the monitoring of traffic, the dispersal of Social Security checks and the ensuring of cybersecurity. Many companies that provide vital technologies for these services process large amounts of data impacting entire societies. Yet the level of transparency required by and applied to democratic governments is not equally applied to the companies behind such services.

Algorithms are not merely the secret sauces that enable technology companies to make profits. They form the bedrock of our entire information ecosystem. Algorithmic processing of data impacts economic and democratic processes, fundamental rights, safety and security. To examine whether principles such as fair competition, nondiscrimination, free speech and access to information are upheld, the proper authorities must have the freedom to look under the algorithmic hood. Self-regulation or ethics frameworks do not make possible independent checks and balances of powerful private systems.

This shift to private and opaque governance that lets company code set standards and regulate essential services is one of the most significant consequences of the increased use of AI systems. Election infrastructure, political debates, health information, traffic flows and natural-disaster warnings are all shaped by companies that are watching and shaping our digital world.

digitization often equals privatization

Because digitization often equals privatization, it means that the outsourcing of governance to technology companies allows them to benefit from access to data while the public bears the cost of failures like breaches or misinformation campaigns.

Technologies and algorithms built for profit, efficiency, competitive advantage or time spent online are not designed to safeguard or strengthen democracy. Their business models have massive privacy, democracy and competition implications but lack matching levels of oversight. In fact, companies actively prevent insight and oversight by invoking trade-secret protections.

Transparency fosters accountability

Increasingly, trade secret protections hide the world's most powerful algorithms and business models. These protections also obscure from public oversight the impacts companies have on the public good or the rule of law. To rebalance, we need new laws. For new evidence-based, democratically passed laws, we need meaningful access to information.

A middle way between publishing the details of a business model for everyone to see and applying oversight to algorithms when outcomes have significant public or societal impacts, can and should be found. Frank Pasquale, author of *The Black Box Society*, sensibly speaks of the concept of qualified transparency, meaning that the levels of scrutiny of algorithms should be determined by the scale of companies processing data and the extent of their impact on the public interest. Failure to address and fix the misuse of trade secret protections for this purpose will lead to the shaping of more and more digitized and automated processes in black boxes.

The level of algorithmic scrutiny should match algorithms' risks to and impacts on individual and collective rights. So, for example, an AI system used by schools that taps and impacts data on children requires specific oversight. An AI element in industrial processes that examines variations in the color of paint is, by contrast, of a different sensitivity. But AI stretches beyond the physical world—into the inner workings of machine learning, neural networks and algorithmic processing.

Some argue it is too early to regulate artificial intelligence or insist that law inevitably stifles innovation. By empowering existing institutions to exert their oversight roles over increasingly AI-driven activities, these institutions can regulate for antitrust, data-protection, net-neutrality, consumers' rights, safety and technical standards as well as other fundamental principles.

The question is not whether AI will be regulated but who sets the rules. Nondemocratic governments are moving quickly to fill legal voids in ways that fortify their national interests. In addition to democratic law-making, governments as major procurers of new technological solutions should be responsible buyers and write public accountability into tenders.

Many agree that lawmakers were too late to regulate online platforms, microtargeting, political ads, data protection, misinformation campaigns and privacy violations. With AI, we have the opportunity to regulate in time. As we saw at Davos, even corporate leaders are calling for rules and guidance from lawmakers. They are coming to appreciate the power of the governance of technologies and how technologies embed values and set standards.

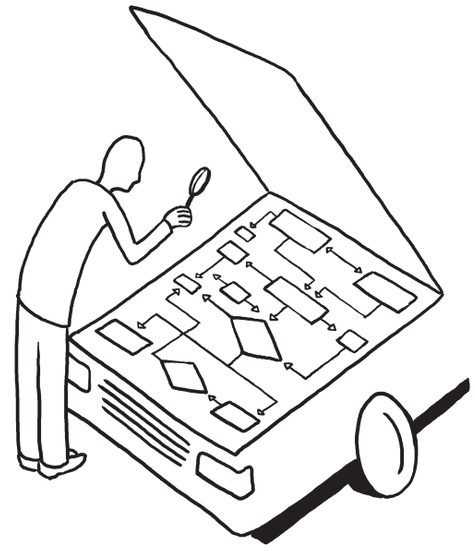
Reaching AI's potential

While much remains to be learned and researched about AI's impact on the world, a few patterns are clear. Digitization often means privatization, and AI will exacerbate that trend. With that comes a redistribution of power and the obscuring of information from the public eye. Already, trade secrets not only shield business secrets from competitors; they also blindside regulators, lawmakers, journalists and law enforcement actors with unexpected outcomes of algorithms based on their hidden instructions. AI's opaque nature and its many new applications create extraordinary urgency to understand how its invisible power impacts society.

Only with qualified access to algorithms can we develop proper AI governance policies. Only with meaningful access to AI information can democratic actors ensure that laws apply equally online as they do offline. Promises of better health-care or the just use of AI in extreme circumstances such as war will reach their potentials only with access to algorithmic information. Without transparency, regulation and accountability are impossible.

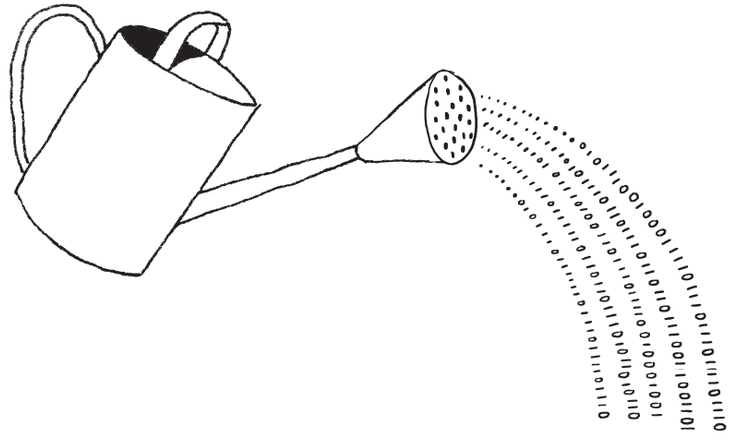
Technology expresses our values. How will we be remembered?

We are at a critical juncture. Our values are coded and embedded into technology applications. Today, companies as well as authoritarian regimes direct the use of technology for good or evil. Will democratic representatives step up and ensure AI's developments respect the rule of law? We can move beyond hope and fear only when independent researchers, regulators and representatives can look under the algorithmic hood.




do we sometimes act differently
what motivates your choices
do you
more valuable than other living things
what makes wrong
do the wrong things
is justice a property of the world
dog moral
are humans kind
would you

taking care



# of business

*about hilary mason & jake porway on pages 87 & 88

The private sector's lens on responsible AI

a conversation:

What does responsible AI, mean to you?

+hilary Responsibility means thinking from the very beginning about potential impacts when building systems. This includes testing as best as possible, understanding that there are often errors and biases not discovered in the development and testing process and having mechanisms to report and correct what may have been missed.

Responsible AI means understanding the use of predictive technology and its impact on people.

This has many layers. It may be allowing a human override when necessary.

Responsible AI is not a technology problem. There is no technical checkbox. You cannot run an ethicize-my-work program and be done.

Responsible AI has to be owned by the product leaders, the business strategists and the people making business-model decisions as much as it is owned by the technologists doing the technical work.

+jake Ultimately, the responsibility for any technology comes down to who has oversight of a system and who says yes or no. It depends on who can say this goes forward or not. It's funny that we automate these processes and tasks and let AIs do their thing.

Do we ignore oversight in any other situation? Like not checking on whom the hiring manager is hiring, for example? We enforce ethical AI by looking at outcomes. We should trust but verify. As engineers, we should be thinking about responsibility of oversight of our systems more that way.

> if a company wants to commit to building responsible AI, it has to commit to building a responsible business that means leadership has to believe that that's the right path forward

+hilary We need to think about how we evolve the practice from one that focuses on the math to optimize the objective function, without regard for impact or testing, to one in which testing is a required step in the development process. You cannot escape human ownership and credibility. Still, there's no test that will solve for this without thinking about who might be impacted in positive and negative ways.

The broad sentiment in the community and at most companies—the people who hire and manage, not just the people with hands on keyboards—is that there's no one process for this.

If a company wants to commit to building responsible AI, it has to commit to building a responsible business. That means leadership has to believe that that's the right path forward.

It doesn't matter how many engineers or technologists ultimately leave their employers, because they can always hire people who share their values. The value system is a big piece.

There is a very broad conversation around excellence; a piece of it is also responsibility. A lot of people are drawn to this work because they care about that piece. So I'm very hopeful for the next ten years.

How are folks in the trenches grappling with these challenges?

+hilary The data science and AI community realizes it has the power to advocate for how they would like to do the work.

That comes from a strong hiring market. Employees can easily move if they don't like what their company does. They can and do. Like the No Tech for ICE movement. People do not want to support something they strongly feel is wrong.

+jake How can we make whistle-blowing something more than a high-risk situation? Because we do see situations where engineers are shuffled out the door for speaking up. How do we make it safe and actionable to push back? Without that, ethical codes or responsibility training won't make a difference.

A bigger question is how much we want engineers to make ethical decisions. I hear a lot of conversations putting the onus on engineers not to do unethical things. But identifying or assessing what's "ethical" isn't always easy. For example, one patient diagnostic system overrecommended oxycodone to make more profits. It was a clear example of an algorithm doing harm so that its company could make more money.

People were up in arms. They were livid that engineers had coded this overprescription feature into the software and did not speak up.

But people assume that distorted outcomes were obvious to the engineer creating this feature. Does that spec come down the line to them as "Let's kill people"? Certainly not. It may be described as a "medication recommendation feature" that senior executives have requested for certain outcomes.

Do you want the engineer to be the one deciding whether the feature prescribes too much oxycodone? How would an engineer know without being a physician? If the engineer pushed back and the feature was removed, would you have the flipside headline, "AI denies pain medications to people in pain"? You may want an engineer to raise questions. But you probably don't want the engineer making the ethical decisions. So, do we have agency to push back as engineers? Do we have context to know when to push back?

+hilary Engineers are building more leverage but do not have agency. And even when they do, they quit their jobs and go somewhere else. And the work continues.

What might illuminate a path forward toward responsible AI?

+jake To talk about responsible AI, we first have to address the goals of a system. AI is basically an accelerator of the values in its system. AI makes reaching the goals of that system faster and cheaper. So, it's important to recognize that responsible AI is impossible without responsible systems. For example, most companies are designed to make money only from their products. So, one version of being responsible focuses on how companies can do less harm and how engineers can be more ethical.

Then there's a version of responsible AI where the technology is applied to a human challenge. For example, what are the apps for getting clean water to people, and what might that look like—responsibly? How can machine-learning or AI help? This takes on a close but slightly different lens, starting less from the AI solution and more from how we can solve this human

problem through an AI intervention so that many more do not die. In that version of responsible AI, we use AI to support systems that have human goals—goals for civil society.

So, when we think about a path toward responsible AI in that second context, we have to ask how we will build that tool and who will build it. There may not be a profit motive. How might we also use our skills to achieve goals like social prosperity? How do you get the technologists on payroll to do that? AI-for-good tools are a cost center. Unfortunately, a lot of social prosperity is a cost center!

However, companies are increasingly putting money and effort behind having their engineers be a part of social-good projects or finding ways to share their data safely for social good. Microsoft has an established AI for X program where X can be Earth, oceans or other social causes. Companies are putting millions of dollars and engineering capabilities and technology into X and partnering with UN agencies or nongovernmental organizations [NGOs] to see how a technology can be applied.

For example, Johnson & Johnson just put $250 million behind digital health workers with a view to exploring with UNICEF and USAID how machine-learning can be used on the front lines of health. Accenture and NetHope are building capacity in the social sector; they have just created the Center for the Digital Nonprofit. And they're investing long term in digitizing the civil society sector and creating responsible digital nonprofits.

Multistakeholder partnerships are clearly the way people and businesses can together make change.

We're involved with a group called data.org, launched by The Rockefeller Foundation and the Mastercard Center for Inclusive Growth, which helps multistakeholder partnerships deepen their social impact through data science. Our work uses AI to boost the effectiveness of the health-care that community health workers provide. Teams are building algorithms that identify which households need care most urgently, and they're using computer vision to digitize handwritten forms to modernize data systems. This kind of innovation is unlikely to come from the private sector alone—or just NGOs. So, we're bringing together folks from different sides of the table to build the AI they want to see in the world. It's a winning approach involving businesses, foundations, NGOs and other actors.

Another strategy is establishing a consortium of companies and agencies thinking about AI safety and responsibility, such as the ABOUT ML group of folks from Microsoft and Google, which wants to create a system to improve the explainability and understanding of the algorithms their companies are building. In both of these examples, companies are devoting their resources to partnerships that allow us to build AI for human prosperity.

+hilary We need to shift the system, get more granular. For example, build a malware-detection model to expose how testing is done.

I'm interested in examples of data scientists' sitting at their computers, doing their jobs and looking at how someone else has done this right.

How and where are companies doing this right?

+hilary Some people are very thoughtful about their image tester bias. They examine almost every real data set for extreme class bias and for how they accommodate classifications.

There are also some things that shouldn't exist. Like the facial-recognition start-up that's selling police departments photos harvested from social media. This violates fair use, copyright and permission terms. Or AI video interviews that are replacing face-to-face ones.

+jake Companies implementing ethics codes in their engineering departments are at least starting the conversation. But we also have to consider how we think about success. Because if we're not clear about what responsible AI looks like, no one company can get it right. In the AI interview case, how would you know if the AI interviewer was biased? This algorithm may be better than the status quo, but it's also encoding a set of consistent biases at scale—which the status quo did not.

Unfortunately, running a rigorous experiment to determine whether your AI interviewer is better or worse than a human is really hard because a huge set of complex economic and demographic factors make it hard to assess such AI systems. This comes back to responsibility. That's why we're having conversations about optimizing for profits and about how these things correlate with numbers of sales. We can all agree on pretty straightforward profit metrics. But incredibly difficult philosophical questions are not easily quantified.

Success in the real world is hard to quantify because the code is too complex and because we all have independent sets of values for how we think the world should be. But AI systems work only when they have very specific objective functions. The greatest trick AI will pull off will not be taking over humanity. Often, we're not explicit enough about what success looks like in society.

> "we're not explicit enough about what success looks like in society

Some big tech companies say they want to be regulated. What is your take on that?

+hilary I'm pretty cynical. When large companies ask to be regulated, they're asking to entrench their advantage in an area that's changing and developing very quickly.

Let's say you put a review process around the deployment of any AI model that costs around a million dollars. So only companies with global scale can afford to use this model.

So small start-ups can't compete. But the Googles on a global scale will invest and try these things. So, I worry deeply that the kinds of regulation these large companies will push through will strongly advantage them and create an oligopoly with access to broad technology and destroy efforts of smaller organizations to use the technology effectively.

On the other hand, regulation could encourage broader innovation through data ownership and data portability—meaning, being legally required to explain when and how you sell data from one organization to another.

+jake When I think about regulation innovation, I think about Kenya. Kenya has a long history of experimentation in the digital space. But the innovators are usually not Kenyan, and collect data that doesn't go back to Kenyans. For example, people were building payment systems that collected Kenyan citizens' data, but creditors sold data they'd just acquired. It was problematic.

So, the Kenyan government passed a modified version of the European Union's General Data Protection Regulation to protect citizens' data. The government also thought the new regulation would bring business in, because it required that data stay in-country. I heard that Amazon is setting up a data center in Kenya—perhaps because of this policy. So, some government regulation can help business and government.

+hilary I'm interested in the California Consumer Privacy Act which became effective in June 2018. If you share data about citizens of California, you're required to disclose it. We'll see what happens.

What about incentives and voluntary or involuntary approaches? What should nonprofits be thinking about?

+hilary I would love to see more collaboration and understanding. People do mess up. Companies are not innately evil. If we give more space to learn and improve over time, I think we'd get to better outcomes. That's really tough right now because even those talking are not doing so in public.

+jake It's really difficult for individuals to change systems. They tend to adopt the values and systems they're in.

We need to think deeply about how complicated and challenging responsible AI is. Mitigation is a common theme. The question is not how do we stop this, but, rather, what should responsible AI look like? We need to be open, considerate, and to reflect on solutions.

There's a nuanced specter of risk. Things feel fairly histrionic whenever AI is perceived to be unethical. Not everyone is the worst offender.

We need to share stories about industries that have regulated well and understand why. Could we not follow the path that is addressing climate change?

We all know that climate change is a huge issue that crosses national boundaries, and yet we're still driving cars. We're not saying Toyota engineers should be rising up and protesting. BP still has plenty of engineers.

And we have other safeguards for the environment. The Environmental Protection Agency as a regulatory body is only as strong as the governments we elect to enforce our laws. We're in a similar spot with AI.

> "companies absolutely have a role to play, and they shouldn't be forced to shoulder this burden alone it comes back to us, as people

Companies absolutely have a role to play, and they shouldn't be forced to shoulder this burden alone. It comes back to us, as people. People with a vision of how AI should be used in their lives should rise up and vote for regulation, vote for tech-literate politicians and find ways to measure AI models that are used in the public interest so we can ensure we're getting what we need. At the end of the day, AI is for us. So, we must define how this stuff works. And we must hold AI accountable. Its level of responsibility must reflect what we as society need.

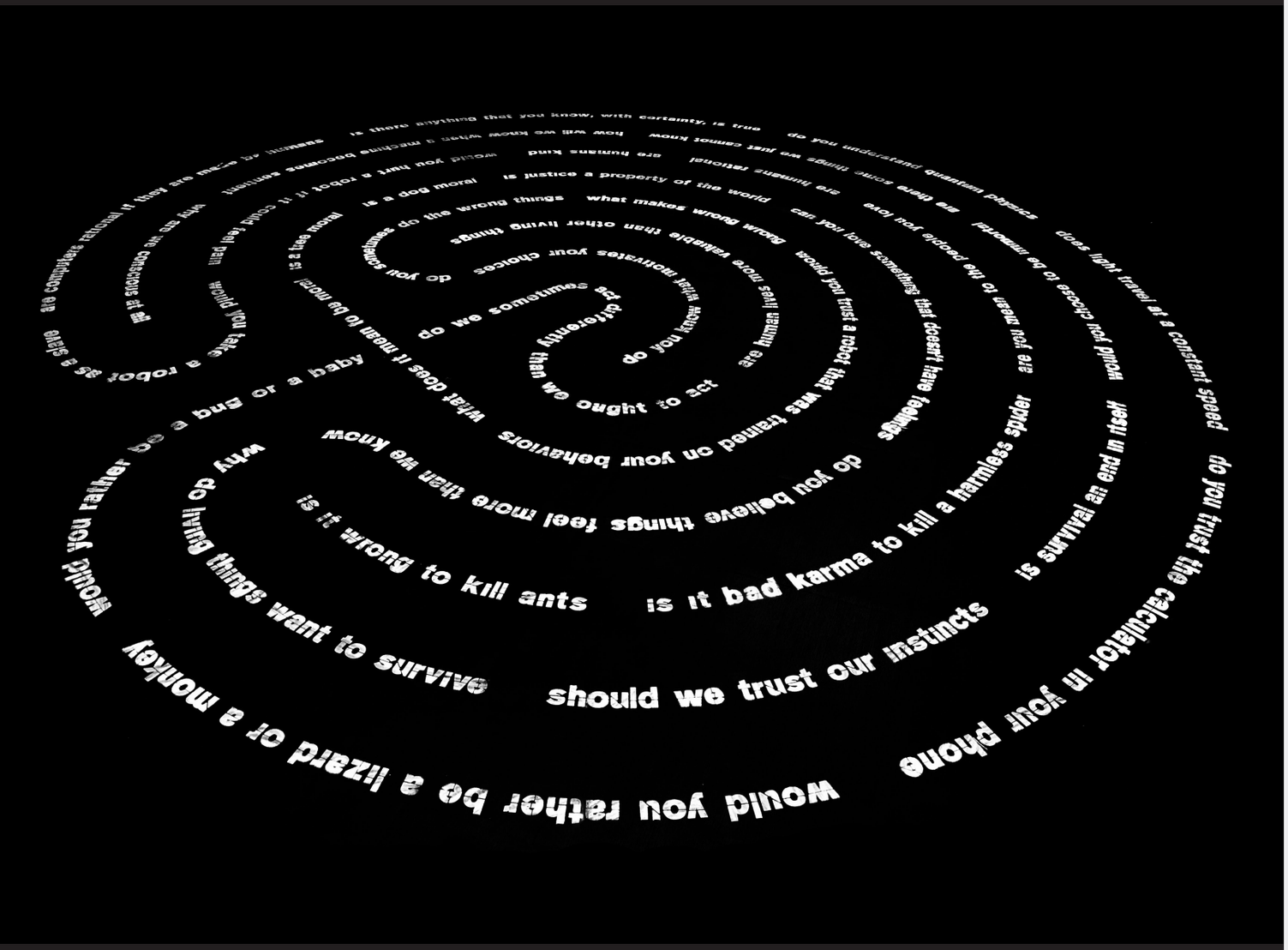
moral labyrinth

- **Would you trust a robot that was trained on your behaviors?**
- **How will we know when a machine becomes sentient?**
- **What does it mean to be moral?**

How can we program values into an intelligent machine when we do not agree on what we value? Values are wildly different across cultures and individuals, and most individuals hold many conflicting values. Beyond that, our behaviors often conflict with our stated values. Should AI systems learn from what we say or what we do?

Moral Labyrinth is a walking labyrinth composed of questions—from the seemingly simple to the slightly absurd—encouraging viewers to examine their own values and assumptions. Can such reflections encourage us to be more honest, humble and compassionate? How will our beliefs, and their incongruities, manifest in our technologies?

The work is a meditation on perennial—and now particularly pressing—aspects of being human.



Walking Labyrinth, dimensions variable, 2018–2020. The version included in this publication was five meters wide and made entirely of baking soda. (Mozfest, Ravensbourne University, London 2018)

making AI work for humans

amir+baradaran & katarzyna+szymielewicz & richard+whitt

The web is light-years away from its inventors' goal of expanding individual freedom and power through new connections and information sources. Understandably. Platforms are reluctant to give up control over our data storage and stories—or our marketing profiles. Users' stickiness—or their propensity to stay on a web page—brings big platforms big profits from advertisers. That concentration of data power also dissuades other developers from competing with large platforms. The result? Platforms have limited accountability, and users have limited alternatives.

It’s broken but can be fixed

How to tilt power and agency back to end-users? New regulations are an obvious and frequently made suggestion. But to date, they are neither fast nor forceful enough. Another common idea is to break up the platforms. But like a game of Whack-a-Mole, over time similar woes will remerge because, as Tim O’Reilly noted in his essay, the same perverse incentives remain. Also under discussion: a utility model much like that used for telecommunications companies, converting platforms to publicly owned companies with data portability—analogous to phone number and contacts portability. All of these approaches strive to increase accountability, which is sorely needed. But their sanctions and obligations primarily deter and punish bad actors.

> we very intentionally use the term humans rather than users because humans are active rather than passive agents

We think there’s a better way. A more human way. A way that doesn’t require dismantling the current system and that expands the awesome reach and potential of AI for good—for all humans. As lawyers who’ve worked in government, at major platforms and at public interest groups—and a technologist/designer who has ripped apart AI’s guts and put it back together—we’ve been in war rooms, legislative chambers and garages. And we strongly believe that a new paradigm is possible. We believe we can restructure a web ecosystem to give more power and agency to end-users and that the web can better serve the interests of a range of stakeholders, not just those chasing profits. Our proposal is in its infancy, but it has many precursors, and we are starting to see its contours.

Humans at the core

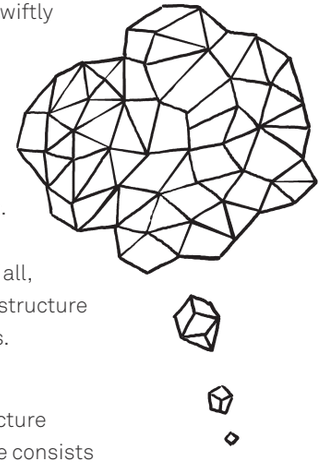
Our new ecosystem puts humans in the driver’s seat and makes possible new business models from trusted platforms through smart interfaces that replace current patterns of data exploitation with real transparency and empowerment. We very intentionally use the term *humans* rather than *users* because humans are active rather than passive agents in these models, providing vital data fuel. In this new ecosystem:

+ **Humans can access, control, verify and shape their personal data and narrative** and its flow without losing contacts, prior posts, asking platforms for permission to keep these things, or additional application downloads. Today, theoretically, humans have a right of data access. But in practice, they receive only a curated picture of their data and cannot do what we are proposing. In the new ecosystem, should humans object to aspects of a profile, those aspects can be deleted in one click. Result: advertisers and commercial developers can access only preapproved information. Humans can thus control their own narratives and their narratives’ journeys. They can also connect to other social networks and access other humans’ data—with permission. Finally, humans outside their social network can follow all public updates generated by humans on this platform in read-only format by using a simple RSS (really simple syndication) function via a standardized API (application programming interface, or basic software instructions).

+ **Other companies and noncommercial developers can train their algorithms or build their own models on existing platform data and stories to offer alternative solutions, functionalities, experiences or privacy and security standards.** Consider a newsfeed that ceases at 10 p.m. or that excludes violent content. Or a newsfeed rigorously fact-checked by an independent news agency. Or a kill-the-newsfeed plug-in. Access to data made possible by a standardized API creates opportunity for such innovation. A maker can ask humans about their expressed preferences and the news they seek (which is impossible today) to propose a curated newsfeed. But the maker may also need limited (read-only) access to statistical models previously developed by Facebook for Facebook’s own purposes. For that, a regulator would need to approve their justification for competitive or public-interest purposes—in an approach akin to eminent-domain rulings.

+ **Public institutions and researchers can use this same information and analysis to help solve hard problems.** By taking a Creative Commons approach, they can aggregate humans’ output data (no longer personal) and statistical models developed by dominant platforms for purposes such as medical research, tracking and responding to pandemics or planning infrastructure upgrades. Possible applications might include matching soup kitchens with supermarkets for food nearing its sell-by date, tracking emissions in moving trucks or analyzing energy consumption patterns in personal devices to calculate their environmental cost. Some of this is already under way. For example, Uber has furnished driver data to the Organisation for Economic Co-operation and Development for tackling gig-economy issues. Google has done the same with US transit agencies for subway system upgrades.

Our new paradigm is a greenfield opportunity for platforms and those in their ecosystem, large and small. Benefits might include new revenue streams and broader visibility from new goods and services; reputational gains from fomenting trust among existing customers and staff (Amazon employees, for example, protested the sale of the company's facial-recognition product for questionable purposes); new business models that may expand reach or reduce costs; better adtech and martech experiences for advertisers and humans; new job categories like ethical data brokers; more-satisfied humans through novel technology options such as personal AI assistants (which we'll go into later); and more control over both humans' and platforms' futures. But time is running out. We must act swiftly in our shared interests.



Rethinking the web's virtual and human infrastructures

A rough outline of how the new ecosystem might work follows. It is achievable with help from experts in multiple domains, including technology, law and business management. Best of all, our vision for a decentralized, open, human-centric web infrastructure builds on existing social networks and commercial databases. Please be in touch with your thoughts and expertise.

Systems and design thinking guide the two types of infrastructure we propose: virtual and human. The new, *virtual* infrastructure consists of interfaces, portals, standards, protocols and other technical means of mediating between existing platforms and humans. The new, *human* infrastructure involves new kinds of organizations and agents that can help humans on the edge of the web access data for their own noncommercial purposes. Three steps make new virtual and human infrastructure creation possible:

+ Breaking up is hard to do: Separating data storage and analysis

To create an open and human-centric data infrastructure, the breaking apart of the collection, transmission and storage of data from its computational analysis is key. That's because many of the statistically significant patterns that emerge from platforms' deep and pervasive access to our data fall on the cutting room floor because the data cannot be easily monetized. So humans lose some of the upside of their stories that could be used for tools and applications that support human flourishing.

In our proposed approach, humans take back control of their algorithmic destinies. Regulations or market forces compel platforms and data brokers to develop a standardized API that enables humans to easily see and verify

in the new space between the storage of data and its interpretation
new roles and new ways of using data for noncommercial purposes can emerge

their data and move it elsewhere—to, for example, a competing social network if desired. Alternatively, the data can be left on the platform, but humans exert far more control over their stories[1] (or, for platforms, their marketing profiles) because data analysis sits elsewhere. An independent regulator such as a consumer protection body will need to define and control the structure of such an API.

In the new space between the storage of data and its interpretation, new roles and new ways of using data for noncommercial purposes can emerge. Other stakeholders—new competitors, researchers or public service providers—can gain access to data inputs, data outputs and models used for algorithmic processing—depending on their needs.

An example of such a human-centric technology platform is Sir Timothy Berners-Lee's Solid project, which shifts and corrals one's data to a local, home-based pod (handheld device) from servers across the world. And Finland's Ministry of Transport and Communications has its MyData project, with a framework, principles and a model whereby individuals can access their medical, transportation, traffic, financial and online data sets in one place, decoupling data storage and its analysis for consent-based release and sharing of data to groups, including public interest research data banks.

Another step in the right direction in the US ACCESS Act, a bipartisan bill introduced in October 2019. The act would grant new authority to humans so they can connect directly with platform networks for interoperability and move their personal data to trusted entities, for portability.

+ The human infrastructure: Trusted stewards who help manage our digital lives

Freeing personal data from its commercial moorings can bring enormous benefits *and* complexity based on novel choices humans can make. But today's ecosystem lacks both human and digital agents to manage the complexities. A human independent third party could shift the current dynamic from one of passive platform-services users to empowered clients.

New and trusted fiduciaries can serve as the *digital life support system* for humans (or clients). That notion builds on the common law of fiduciary obligations, which includes the duties of care, of confidentiality and of loyalty. Humans can select their own digital fiduciary to act on their behalf. Typical functions might include:

+ *Basic client protection such as managing passwords, updating software and establishing privacy settings*
+ *Filtering the client's personal data-flows to reflect personal interests*
+ *Using advanced-technology tools, such as a personal AI, to promote the client's agency and autonomy*

To wrap your head around this intermediary approach, consider a residential real estate analogy. The seller's agent ostensibly works for both parties to secure a deal. But in truth the agent is working only in the seller's interests. The buyer's agent, by contrast, serves solely the buyer.

In our ecosystem, the seller's agent is a platform company; the buyer's (here, the human's) agent is a trusted intermediary. In other industries, similar separate agents that serve different interests within the same transaction include (1) a pharmaceuticals manufacturer (selling drugs) and a local pharmacist (tasked with giving sound advice) and (2) a bookseller (selling books) and a librarian (tasked with giving sound advice and protecting patron privacy).

On this point, the ACCESS Act legislation mentioned earlier would let humans delegate their interoperability and portability rights to a trusted third party—a custodian—operating under strong fiduciary duties.

Some digital third-party precursors already exist. Digi.me enables humans to download data from various sources to their phones in encrypted form, which other services can then process on-device.

think of a personal AI assistant as a virtual version of Alexa which is 100% on your side

+ **The virtual infrastructure: An Alexa as your personal-data police force made possible by separating the computational layer**

A longer-term but increasingly viable evolution of the API is the personal AI assistant. Think of a personal AI assistant as a virtual version of Alexa which is 100% on your side. This augmented reality avatar acts almost like a librarian—fetching *data* on demand instead of *books*—and protects your logs from tracing. Your personal AI assistant takes on those otherwise burdensome tasks, which helps separate personal data from the stories that result from its analysis. Building APIs into the platform companies' computational systems lets humans and their digital agents control their data-flows and analyze data and patterns for their own needs.

Sometimes referred to as *on-device, off-cloud AI,* these applications hold enormous potential to represent humans in daily interactions with the web. Among tasks that a personal AI assistant can perform are protecting a human's online security from rogue actors or hackers, projecting a human's term of service to websites (rather than the reverse) and the bidirectional filtering of newsfeeds, social interactions and other content-flows on the web. A personal AI assistant can even challenge the efficacy of algorithmic systems—representing the human in, say, disputes with financial, health-care and law enforcement entities—for bias, error and other flaws with potentially serious consequences. Finally, it can query, correct, negotiate and demand that the client be left alone.

This virtual zone of trust and accountability can stretch into the offline space. Today, companies and governments are embedding billions of sensors in smart speakers, microphones, cameras and wearables to extract and act on our personal data, including information on our location, facial expressions, or physical state. (In 2016, 325 million wearables were already in circulation.) A personal AI assistant can actively prevent these devices from unapproved surveillance and extraction of data. Instead, it blocks signals or negotiates with the sensor provider, fortifying agency that would otherwise not be possible.

This concept is gaining traction. Stanford University is working on a virtual assistant called Almond, which retains a human's personal information, thus reducing dependence on big platforms for services. US engineering trade association, the Institute of Electrical and Electronics Engineers (IEEE), recommends a proxy, or "trusted services that can… act on your behalf… a proactive algorithmic tool honoring their terms and conditions." In fact, work is already under way at the IEEE to develop standards for personal AI assistants through its P7006 working group. Precursors that can help build-in interoperability and portability include US Federal

Communications Commission concepts developed in the 1980s and 1990s to encourage telecommunications competition.

A standardized API, perhaps evolving into personal AI, is the first step toward an interconnected infrastructure controlled by citizen-humans, not advertisers.

Democratizing AI: Serving humans on the edge

The problems related to today's commercialized data ecosystem are well-known and will evolve and grow. Clearly, an ecosystem based on the wrong premise is not sustainable, and a radical change is needed. Infrastructure, services and interfaces that are more responsible and human centered are required. Our most urgent task is to translate these values into tangible, practical solutions.

Our proposed ecosystem shifts the power from platforms and advertisers at the core of the web to humans and other stakeholders at the edge of the network. Result: a more sustainable, human-centric ecosystem in which people can reclaim control over their data and their digital lives. And not just data stored or generated about them, but, more importantly, control over how interpretation and application of their data influence their and their peers' life chances and life choices.

> our data can be used to deliver social value and ethical services that are free from commercial ends and hidden influence

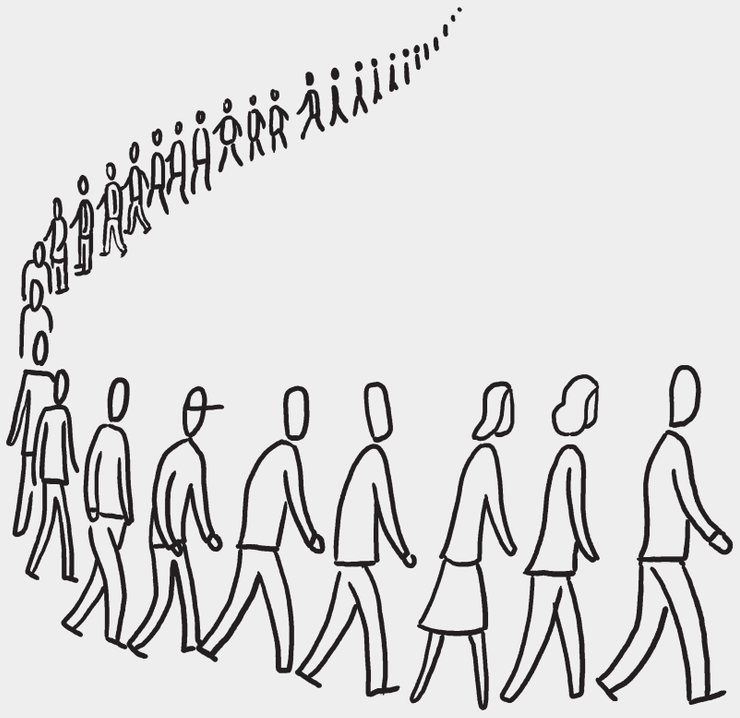
This paradigm shift also opens the online world to a new set of new actors and roles, often pursuing loftier goals than transactions or profits. Stakeholders might include nonprofits, B corporations, public service providers and data trusts. Delivering value without a service—such as by targeting societal problems in a nonprofit or public model and a service not-in-the-bundle with influence—is also possible.

By keeping and aggregating our personal data under our control, we can solve our own challenges and tell our own stories—on our own terms—and reach others with what we need. Our data can be used to deliver social value and ethical services that are free from commercial ends and hidden influence. Such options are difficult if not impossible in a wholly privatized, profit-driven ecosystem. But they can emerge in an ecosystem redesigned to serve our self-defined individual and societal purposes.

There are of course challenges. They include establishing communications protocols, API structures and new design (user-friendly interfaces; rules-based settings) and governance structures such as data trusts. But those obstacles are surmountable. When complete, this virtual infrastructure will transition power over data from centralized platforms back to humans. Such solutions are not mature enough to be implemented nor even prototyped. But we must begin. Please join us in this endeavor.

1. A Baradaran. Decolonizing "Artificial" Art Making: The Impact of Artificial Intelligence on the Art Ecosystem. Accelerator Series: Technology, Visual Culture, and the Politics of Representation, ed. I Brielmaier. Saratoga Springs, NY: Skidmore College, 2020; Frances Young Tang Teaching Museum and Art Gallery.

data projects' secret to success is not in the algorithm



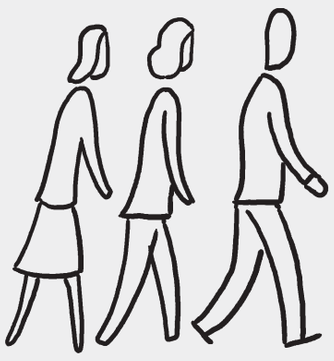
Case study: Digital health services in Burkina Faso

The Data4Good movement often references the sophisticated use of data in the private sector as what the nonprofit space needs—setting aside concerns about ethical corporate-data use for a moment. However, nonprofits face fundamentally different data challenges beyond the obvious requirement of modeling the highest standards of responsible data use.

Companies with business models that rely on the use of data (almost all of them today) employ hundreds of software engineers, designers and marketers to manage, store, analyze and realize revenue from data and thus must continually improve their data management to become evermore sophisticated and stay ahead of the competition.

Success in the nonprofit space, by contrast, generally hinges on the quality of an implementation—not on the cutting-edge algorithm. And as a result, data projects that strive to solve societal challenges at scale are difficult to execute successfully because:

+ *They require a multistakeholder approach involving the individuals or communities facing the challenge, as well as several layers of government and other data holders.*
+ *Their governance is complex because many governments have not yet established privacy and data ownership regulations.*
+ *Better information doesn't necessarily lead to better action. Just as a good policy brief doesn't necessarily make for better policies, better data insights may not necessarily trigger the desired action.*



claudia+juech
founding ceo of the cloudera foundation

Building on strong relationships

The work of the Cloudera Foundation's grantee Terre des hommes (Tdh), Switzerland's leading child relief organization, illustrates these challenges and how they can be overcome.

Tdh's Integrated eDiagnostic Approach supports health-care workers in rural clinics in Burkina Faso in their diagnoses and treatments of preventable diseases in children younger than 5 years old. The health-care workers manage cases through a tablet app version of WHO's and the United Nations Children's Fund's Integrated Management of Childhood Illness (IMCI) protocol. IMCI consists of a sequence of steps and information on how to address common though potentially life-threatening childhood conditions such as pneumonia, diarrhea and measles.

Roughly 1,150 clinics, or 70% of all those in the country, now use the tool for consultations, which topped 5 million by January 2020.

Getting there wasn't easy. Tdh had neither hard evidence of the health benefits nor of the exact cost savings it might gain with the new approach. The organization had spent many years convincing the government to support the delivery of those services, winning approval in 2011 to begin in 39 health centers.

Providing credibility to build a case was Tdh's 40-year history of working in Burkina Faso and its positive relationships with decision-makers after all those years. But what perhaps helped most was co-designing the tool with its future users at health facilities and the district level from the very start.

Since the project's inception, 11 iterations have incorporated feedback and suggestions for improvement from all user groups. Tdh's clear commitment to transfer the data and its management to the government was another critical success factor, as was holding workshops to share learnings and facilitate the handover since the tool's launch.

Incentives matter

Acknowledging and incorporating incentives into the design of the digital initiative also played critical roles in ensuring the tool's buy-in and use. And even though motivations to use or support the app may have overlapped in some cases, all stakeholders cited distinct benefits that were subsequently prioritized for design and execution by:

+ ***The Ministry of Health (MoH):*** *Cost savings, better quality of care and disease surveillance and more timely and accurate information*
+ ***Those at the district level:*** *Improved workload management and performance management*
+ ***Frontline health-care workers:*** *Easier ways to work because MoH support meant that workers would no longer be required to enter data twice —into both paper and digital versions— and elevated professional status stemming from the use of a hand-held device rather than a paper-based version*
+ ***Parents or guardians:*** *Trust in technology that would lead to higher quality of care, such as a more thorough check of all of a child's symptoms*

So strong were incentives in some districts where the approach had not yet been implemented that some communities bought the tablets with their own money.

Overwhelmingly, Tdh emphasized in its framing of the new tool the tool's intent to assist and augment human decision-making rather than replace it.

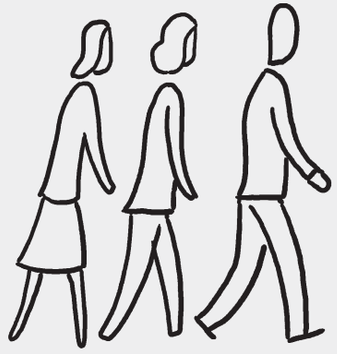
Ongoing exchanges

Successful implementation beyond launch requires a feedback loop that demonstrates to frontline health-care workers the added value of a digital solution. Monthly reports on performance as well as descriptive statistics made possible by the tool, now form the basis for open group discussions in which health-care workers and district heads exchange learnings about what is working and what needs to improve and why, thereby creating a culture of positive, collective pressure to excel.

Also ensuring success were efforts not to fall into the build-it-and-they-will-come trap. For example, before revamping a dashboard visualization, Tdh held a workshop with government officials to jointly determine desired indicators.

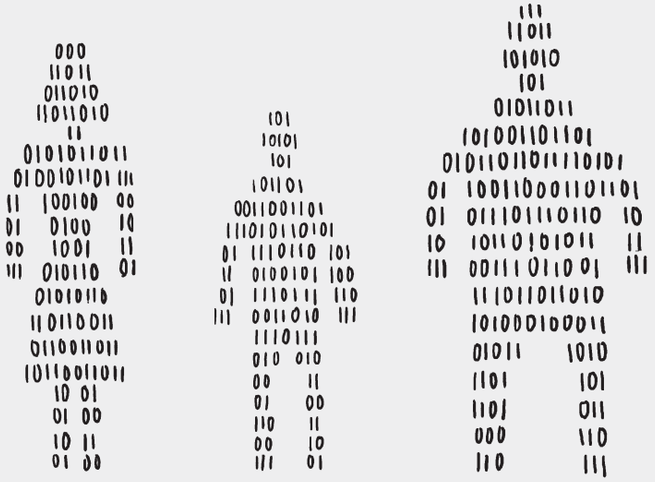
Using data to successfully predict epidemiological seasonal disease trends will require combining the data with other data sets that influence the community, such as weather observations, population movements, transient market events and locations. To obtain such data from other ministries in Burkina Faso, Tdh has added a government data liaison to the local team.

This case study demonstrates that there is no secret recipe to successfully manage stakeholders in a large-scale data initiative. Responsibility and success are two sides of the same coin. Authentic, transparent and frequent interactions; active listening; and responding to the needs—and wishes—of different groups must be embedded in ongoing execution. A successful approach puts humans and all stakeholders at the center, with the digital technologies as catalysts for impact. For groups behind initiatives like these—nonprofits and funders alike—investing in building and nurturing relationships is paramount.



With thanks to Thierry Agagliate, head of innovation at Terre des hommes Lausanne, and Riccardo Lampariello, head of health program at Terre des hommes Geneva, for their input and review.

inclusive AI needs inclusive data standards



Designing AI for the public good is in our hands

Modern artificial intelligence (AI) was hailed as bringing about the "end of theory." To generate insights and actions, no longer would we need to structure the questions we ask of data. Rather, with enough data and smart enough algorithms, patterns would emerge. In this world, trained AI models would give the "right" outcomes—even if we didn't understand how they did it.

Today that theory-free approach to AI is under attack. Scholars have called out the bias-in, bias-out problem of machine-learning systems, showing that biased data sets create biased models and, by extension, biased predictions. That's why policy-makers now demand that if AI systems are used for making public decisions, their models must be explainable by offering justifications for the predictions they make, but a deeper problem rarely gets addressed. It is not just the selection of training data or the design of algorithms that embeds bias and fails to represent the world we want to live in. The underlying data structures and infrastructures on which AI is founded were rarely built with AI uses in mind, and the data standards—or lack thereof—used by those data sets place hard limits on what AI can deliver.

> the underlying data structures and infrastructures on which AI is founded were rarely built with AI uses in mind

Questionable assumptions

From form fields for gender that offer only a binary choice, to disagreements over whether or not a company's registration number should be a required field on an application form for a government contract, data standards define the information that will be available to machine-learning systems. They set in stone certain hidden assumptions and taken-for-granted categories that make possible certain conclusions—while ruling others out—before the algorithm even runs. Data standards tell you what to record and how to represent it. They embody particular worldviews. And they shape the data that shapes decisions.

For corporations planning to use machine-learning models with their own data, creating a new data field or adapting available data to feed the model may be relatively easy. But for the public good, uses of AI—which frequently draw on data from many independent agencies, individuals or sectors—the syncing of data structures is a challenging task.

Opening up AI infrastructure

There is hope, however. A number of open-data-standards projects have launched since 2010. They include the International Aid Transparency Initiative, which works with international aid donors to encourage them to publish project information in a common structure, and HXL, the Humanitarian eXchange Language, which offers a lightweight approach for structuring spreadsheets with who, what and where information from different agencies engaged in disaster response activities.

When those standards work well, they enable a broad community to share data that represents their own realities, and they make the data interoperable with data from others. But for that interoperability to happen, standards must be designed with broad participation so that they avoid design choices that embed problematic cultural assumptions that create unequal power dynamics or that strike the wrong balance between comprehensive representation of the world and simple data preparation. Without the right balance, certain populations might drop out of the data-sharing process altogether.

To use AI for the public good, we have to focus on the data substrata on which AI systems are built. That focus requires primary focus on data standards and far-more-inclusive standards development processes. Even if machine-learning lets us ask questions of data in new ways, we cannot shirk our responsibility to consciously design data infrastructures that make possible both meaningful and socially just answers.

unlocking AI's potential for good

requires new roles and public–private partnership models

Data collaboratives and chief data stewards can help forge alliances and scale AI's use for higher purposes

Most observations of our data era's shortcomings focus on the misuse of data. But our failure to capture and apply existing data for the public good is a glaring gap in our discourse.

Indeed, abundant opportunities to reuse data for benevolent ends lies in, for instance, call and online purchase records, as well as in sensor and social media data that is increasingly used for AI applications. In most cases to date, these types of data are stored and controlled by companies. But functional and responsible access to such timely and comprehensive data sets can help public agencies and researchers develop algorithms that transform how we make decisions and solve public problems. If designed properly, AI can also help public service providers better target those who need services, and it can reduce costs over time. Better insights from better data help governments better understand what works and where those in real need are, thereby cutting waste and avoiding misuse of public services.

To capture and harness the power of data and AI to improve people's lives, we need to understand and find ways to unlock and responsibly reuse private data for the public good through new types of partnerships and a dedicated profession that would initiate and manage such alliances. So important is this sort of work that it calls for a new C-level executive who would report regularly to a company's CEO and board, just as chief operating officers or chief financial officers do today. We call this new role, *chief data steward*. Collectively, these new officers would share and scale their efforts through new alliances we call data collaboratives.

we need to understand and find ways to unlock and responsibly reuse private data for the public good through new types of partnerships

Data collaboratives: public–private partnerships for the data age

Globally, the work of data collaboratives is growing in scope, scale and number. Data collaboratives have emerged from the ashes of disasters like earthquakes in

Nepal and the spread of the coronavirus globally by sheer force of will and the fight to save lives. Now they are branching out beyond urgent, crisis situations. Data collaboratives have used, for instance, call-detail records to track mobility choices and trajectories among women and girls in Latin America to shore up programs that ensure safe transit. Facebook social media feeds are now shared with researchers to analyze the effect of social media on democracy and elections.

Data collaboratives draw on broadly dispersed data and expertise from government agencies, businesses, nonprofits, community groups and activists. Data collaboratives' potential to improve people's lives and strengthen democratic institutions stems from the variety and velocity of the data collected. But such data troves are often scattered within and across organizations and are poorly managed. This disconnect causes tremendous inefficiencies, increased risk of unauthorized access to personal data and lost potential. Data collaboratives, when designed responsibly, can weave together otherwise siloed data and a range of expertise to match data supply with demand in a fair and transparent manner. That meshing would ensure that relevant institutions and individuals can use and analyze data in responsible ways that make possible new, novel social solutions.

Consider today's data collaborative shoots: They are in Santiago de Chile, where the Chilean government, nongovernmental organizations, UN agencies, a telecommunications operator and a university collaborated to analyze how cities are designed in ways that affect women and girls differently—for example, in the locations and availabilities of vital services like health-care clinics, schools, transportation, workplaces and areas that improve well-being such as parks. They are in West Africa, where humanitarian agencies quickly responded to new Ebola outbreaks because NetHope used data provided by the private, public and humanitarian sectors to map the disease's trajectory. And they are global. A Google-Oceana-SkyTruth tie-up helped curb illegal fishing by tracking sea vessels' movements and actions through satellite imagery. These early efforts suggest that companies are ready to share anonymous data for vital services. Imagine the potential of this structure in public health, education, energy, economic development or environmental ecosystems. With scale, a vision and leadership, great things are possible.

But today, setting up such data collaboratives is often prohibitively costly. There are far too few precursors and champions to oversee their design or execution. To take on this herculean, complex task in a private–public context, businesses should appoint and empower senior executives or teams to identify and implement opportunities to unlock the public value of private data. We call individuals who play these roles, chief data stewards.

Chief data stewards

As society has evolved and new needs have emerged, businesses have consistently added new C-level titles such as chief innovation officer and chief sustainability officer. The need for transcompany and sectoral collaborations around data and its use for social good now warrants its own executive role. Anointing and charging these chiefs as cross-sectoral data collaborative leads can fast-track the discovery and pairing of problems that would benefit from higher-quality, more current or more comprehensive data—and from responsible access to the data vaults that contain these treasures. Classifying and cataloging organizations' data assets can bolster the efficacy and speed of the joint work of such alliances for more systematic, sustainable and responsible decisions. Systematically assessing the risks—and the risks of not providing access to data—also makes possible the more responsible use of important data assets.

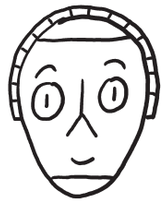
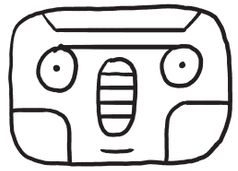
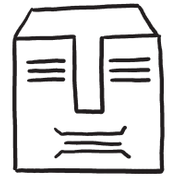
By partnering with public-sector experts and researchers who desperately need—and can wring value from—such data, along with independent intermediaries and other ecosystem enablers, data stewards can lay the operational groundwork for effective collaboration. They can systematize, streamline and accelerate functional access to data for the public interest while ensuring alignment with business, corporate responsibility and societal priorities. As data scientists with links to government, they can also translate insights generated toward action and improved decision-making.

In short, data stewards can unlock the positive potential of our data age and accelerate functional access to private data. They represent an essential new link in the human–data collaboration chain.

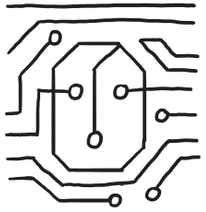
Exactly what should such data stewards be doing, you may ask. To date, their remit elicits considerable confusion based on the misperception that, say, chief privacy, chief data or chief security officers may take on their core work. The data steward's role is somewhat broader. Although ensuring that data remains secure and its privacy protected is part of the work of trusted and effective data collaboratives, a data steward's real mission is harnessing private data for pressing social goals—while preventing harm.

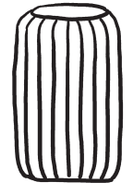
Solving today's intertwined problems demands new ways to develop solutions. Leveraging data and AI for decision making and increased collaboration across sectors can unleash social innovation. Data collaboratives can solve day-to-day problems and respond to cataclysmic crises. Yet to ensure that these data collaboratives are systematic, sustainable and responsible, we need a new human infrastructure. Chief data stewards can champion private data for public-good purposes and channel its use to organizations that unlock its value to help solve pressing global problems. They are the missing keys to generate insights and solutions from data and AI that can transform our world in close partnership with those who can act and will be impacted.

> leveraging data and AI for decision-making and increased collaboration across sectors can unleash social innovation

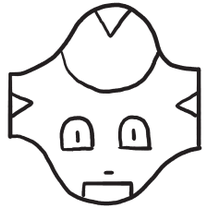




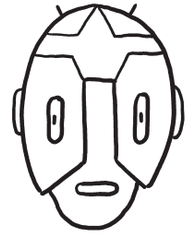
From Ex Machina to Black Mirror,



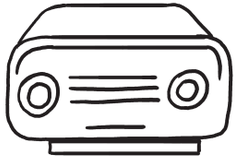
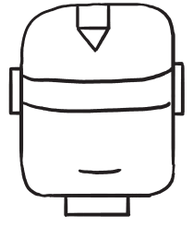


metaphors color our thinking about AI.





How can we use it to our benefit?


making sense of the unknown

We all know what artificial intelligence (AI) looks like, right? Like HAL 9000, in *2001: A Space Odyssey*—a disembodied machine that turns on its "master." Less fatal but more eerie AI is Samantha in the movie *Her*. She's an empathetic, sensitive and sultry-voiced girlfriend without a body—until she surprises with thousands of other boyfriends. Or perhaps AI blends the two, as an unholy love child of Hal and Samantha brought to "life" as the humanoid robot Ava in *Ex Machina*. Ava kills her creator to flee toward an uncertain freedom.

These images are a big departure from their benevolent precursors of more than half a century ago. In 1967, as a poet in residence at Caltech, Richard Brautigan imagined wandering through a techno-utopia, "a cybernetic forest / filled with pines and electronics / where deer stroll peacefully / past computers / as if they were flowers / with spinning blossoms." In this post-naturalistic world, humans are "watched over / by machines of loving grace." Brautigan's poem painted a metaphorically expressed anticipatory mythology—a gleefully optimistic vision of the impact that the artificially intelligent products California's emerging computer industry would make on the world.

But Brautigan's poem captured only a small subset of the range of metaphors that over time have emerged to make sense of the radical promise—or is it a threat?—of artificial intelligence. Many other metaphors would later arrive not just from the birthplace of the computer industry. They jostled and competed to make sense of the profound possibilities that AI promised.

Today, those in the AI industry and the journalists covering it often cite cultural narratives, as do policy-makers grappling with how to regulate, restrict, or otherwise guide the industry. The tales range from ongoing invocations of Isaac Asimov's *Three Laws of Robotics* from his short story collection *I Robot* (about machine ethics) to the Netflix series *Black Mirror,* which is now shorthand for our lives in a datafied dystopia.

Outside Silicon Valley and Hollywood, writers, artists and policy-makers use different metaphors to describe what AI does and means. How will this vivid imagery shape the ways that human moving parts in AI orient themselves toward this emerging set of technologies?

More like Munich or more like Korea?

When we encounter a novel situation that defies established concepts, to make sense of the unknown we tend to search for analogies to familiar past situations. In other words, metaphors "tame" the new. They open it up to the imagination.

Famously, the debate about what to do about Vietnam in 1965 in Lyndon Johnson's presidential administration was ultimately a dispute between those who described the situation as "more like Munich"—thus demanding escalation rather than peace-making—or "more like Korea"—a quagmire to be avoided.[1] In fact, the truth lay in between, not as a blend of previous episodes. Both metaphors were misleading in different ways, and yet they were used extensively in debates and decision-making.

Indeed, as Richard Neustadt and Ernest May argued in their seminal *Thinking in Time,*[2] which offers a critical view of how policy-makers can best make use of history as a guide for both analysis and action, when faced with a novel challenge human nature inevitably leads us to analogies. We must thus both be aware of our implicit analogical frames *and* be explicit about such thinking by naming directly how the current situation *is* both and, perhaps more important, how it *is not* like the historical analogy one references.

metaphors "tame" the new
they open it up to the imagination

We don't only turn to metaphors when confronting the new. In fact, metaphors are vital thinking tools. As German philosopher Hans Blumenberg argued, because there can never be any direct, unmediated access to reality, metaphors are the irreducible lenses through which thought happens. Metaphors shape our knowledge of the world, doing what Blumenberg calls *thought work,* simplifying and thus making accessible complex concepts that without these comparisons we would fail to grasp.[3]

AI technologies challenge
the notion of the human itself

When it comes to AI, metaphors abound because technology makes possible radically new forms of sense-making, perception, feeling and cognition co-produced through people's interactions with technical affordances and infrastructures.

What makes AI so new and unsettling? Unlike many post-industrial-revolution information and communication technologies, AI technologies challenge the notion of the human itself. Philosopher Tobias Rees writes that for people who believe that the human is distinct from and superior to the natural world, to animals and to Earth itself, and that they are more than *mere* machines, to suddenly face machines that appear lively, seductive, competent and clever could be annihilation. AI is like a cracked mirror, reflecting long-held notions of ourselves as humans *and* as humans living with, through and in, complicated conjunction with machines. That colliding image is long overdue for an update.

What are our AI metaphors?

Using metaphors helps nonexperts understand how we build, interact with and regulate technology. "Information just wants to be free," "Data is the new oil" and computer security described as an infection or as a transgression by diseased or "foreign" bodies are some enduring metaphors.[4]

Metaphors often subsume and disguise the creation of technologies. For example, researchers Cornelius Puschmann and Jean Burgess found that the metaphors used for describing big data tend to obscure the material and political conditions of the production and ownership of that data.[5] Through the use of a highly specific set of terms, the role of data as a valued commodity is effectively inscribed (e.g., "the new oil")—most often by suggesting physicality, immutability, context independence and intrinsic worth.

Sometimes AI metaphors are explicit: We need "a Food and Drug Administration [FDA] for AI" or "a Motion Picture Association of America [MPAA] for AI." Some people suggest AI represents a "new kind of market," whereas others liken it to "a human rights challenge." As for AI's regulatory approaches, suggestions include "a peer-reviewed process" for algorithms, "an Institutional Review Board for data uses" or even "a Supreme Court" for algorithms.

Just as the lead-up to the Vietnam war differed from Munich or Korea, so the emergent reality of AI is more complex than any one metaphor suggests. Although we cannot help but think in metaphors about the technology, we need to be aware of how we are using those analogies to describe AI and the ways they emphasize certain features, obscure others and may distract us from what is truly novel about AI. Thus, for example, the metaphor of "an FDA for AI" suggests a government-defined regulatory approach that entails a systematic, defined process for inspecting and approving AI applications. By contrast, "an MPAA for AI" implies self-regulation.

What both of these metaphors hide is that AI is neither a distinct industry nor even a specific set of products. It is a general-purpose technology that has become embedded in myriad products and that is steadily penetrating and rapidly transforming every corner of the economy. A clearer and more explicit assessment of the metaphors used for governing AI should therefore reveal the limits of those metaphors for analysts to be more self-aware of the assumptions—even biases that these metaphors bring along.

AI and culture

Visions of AI are invariably linked to culture. In her collage *Faith and Trust* (2019), for example, artist Şerife Wong examines the dominant metaphors emerging through search results for the term *AI*. By collecting images to identify connections—and gaps—in the words and images we use to describe AI, Wong finds:

There is a preponderance of images that present AI as God, a divine creative force or as a spark—a connection between the human and the divine. Often there is a spark of light—electrical and like sunlight but also evocative of a Frankenstein-like creation. Those images of generative power all draw from Michelangelo's *Creation of Adam* in that technology is a savior of mankind or can give us power over nature.

Wong points to images of robot and human shaking hands; sometimes their outstretched hands are on bodies wearing suits. In the images, AI is our partner, almost human, something to work with and toward. But that "AI is good" narrative creates a duality because it also implies that

there is a preponderance of images that present AI as, variously, god, a creative force or divine or as a spark—a connection between the human and the divine

"AI is (also) evil." The collage's primary colors, Wong notes, are mostly cool tones of blues, greens and purples that convey the feeling of AI as cold. The tones suggest that we perceive AI to be nonhuman—not natural, warm or friendly. And herein lie contradictions: AI is creative, God-like, almost human but also unnatural, cold and not human. The real meaning of AI will emerge from these contradictions.

To echo Rees, AI unsettles our notion of ourselves as human and, by extension, our relationship to things other than human. That said, the range of AI images and AI metaphors is not geographically uniform. In fact, different metaphors about AI prevail in different countries. Americans discuss AI with metaphors that are materially different from those the Chinese or the Indians or the Europeans use. In China, AI is seen as less of a threat, less of a Frankensteinian monster or an engine of oppression the way it is often viewed in the West. In China, AI is more of a symbol of how the country as co-leader in AI development alongside the United States is regaining its rightful place on the world stage.

In Japan, AI-powered robots aren't considered either vaguely or explicitly sinister. They are more often viewed as friendly, as souped-up versions of the famous Tamagotchi hand-held digital pets.[6] Such personification of machines is not new. Buddhists administered funeral rites for the "dead" bionic pet AIBO in 2006, when Sony discontinued its animatronic dog after seven years of popularity. Since AIBO's 2018 relaunch, AIBO owners of all ages gather at cafés marked with its logo. Japan is also the home of a range of other companion robots like LOVOT and PARO. These robots, modeled on a baby harp seal, are often used in medical care and mental health settings, with older people, and with children who have autism.

Against that backdrop, Japanese robotics scientists building *kansei* robots (loosely translated as *affective computing in robotics*) embed into their designs the concept of relationality between the robot and the human. They seek to imbue their robots with *kokoro*—Japanese for the *integration of emotion, intelligence and intention*. It is also the origin of intelligence

and emotion. That unique framing of robots—which is distinct from Western perceptions of robots as not quite human—has helped Japanese roboticists create a national context and a potential market for robotics in Japan and other East Asian countries.

A particular set of cultural norms shapes the contours and limits of human experience and feeling in these metaphors. Robotic traits might include amusement, play and curiosity of a deeply intimate kind with nonhumans. Such traits and attachments could be considered similar to those we ascribe to and form about our own pets. But dogs are not connected to the cloud and don't read our social media feeds. (Cats, however, might!) However, scholar Kate Devlin finds that when it comes to sex robots and fembots, we revert to traditional notions of gender, bodies and sexuality. Gone are the uncertainties related to nonhuman others. On the other end of the spectrum, some Japanese men warm to more conservative traditional holographic "girlfriend"/digital assistants like Azuma Hikari. The assistants embody subservient female attributes through their actions, such as turning on lights before their owners return to empty apartments, ordering their owners' favorite dinners and welcoming orders from their owners.

knowing the range of global metaphors used across the world to make sense of AI will also be valuable for technologists developing AI applications

limiting our engineers' imagining of what these technologies can do and mean to their own norms and habits will limit AI's possibilities

Why consider metaphors?

Lessons from the global AI metaphors landscape stretch well beyond cultural insights to material implications for policy-makers—especially those negotiating global pacts to regulate AI. Because the metaphors bring assumptions about what AI can or may do today and in the future, they shape the debate about how national governments should regulate AI. These distinct beliefs about the role policy-makers in national governance of AI in turn inform—albeit often in inexplicit ways—the ambitions and boundaries that different governments consider for AI transnational regulations and treaty obligations. Understanding different AI metaphors is thus a precondition for understanding and agreeing on global AI regulations.

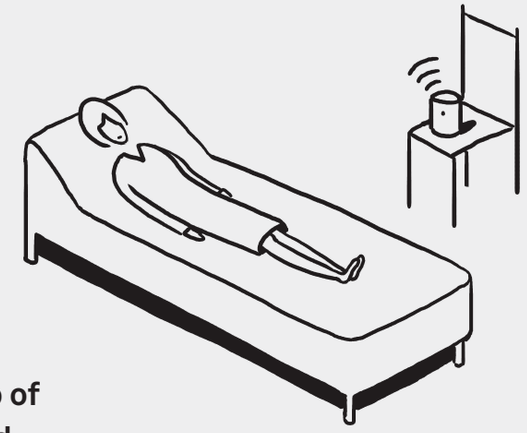
Knowing the range of global metaphors used across the world to make sense of AI will also be valuable for technologists developing AI applications. Limiting our engineers' imagining of what these technologies can do and mean to their own norms and habits will limit AI's possibilities. Cataloging and assessing metaphors used to describe and imagine AI's potential and prospects will help transfer color from these metaphors to creations, thus expanding these technologies' unique and unprecedented possibilities for their more human or multidimensional qualities.

More ambitiously, because AI calls into question our long-standing understanding of the human, assessment of the metaphoric foundations of the discourse around AI will help us imagine our own humanity in radically new ways. These metaphors comfort us in the face of what we cannot yet conceptually grasp. As we grow our awareness of the unique affordances of AI, we may eventually develop new, more adequate concepts about ourselves. On that journey, AI metaphors help us navigate the unknown.

1. YF Khong. *Analogies at War: Korea, Munich, Dien Bien Phu, and the Vietnam Decisions of 1965.* Princeton, NJ: Princeton University Press, 1992.
2. RE Neustadt, ER May. *Thinking in Time: The Uses of History for Decision Makers.* New York: Simon & Schuster, 2011 [1986].
3. H Blumenberg. *Paradigmen zu einer Metaphorologie.* Berlin: Suhrkamp, 1999 [1960].
4. S Helmreich. Flexible infections: computer viruses, human bodies, nation-states, evolutionary capitalism. *Sci Tech Hum Values* 2000;25L4:472-491.
5. Cornelius Puschmann and Jean Burgess, "Big Data, Big Questions: Metaphors of Big Data," *International Journal of Communication* 8 (2014): 20. In reference to Perry Rotella, "Is Data the New Oil?" *Forbes,* 2 April 2012.
6. S Šabanović. "Inventing Japan's "robotics culture": the repeated assembly of science, technology, and culture in social robotics." *Soc Stud Sci* 2014;44:342-367.

complete machine autonomy?

it's just a fantasy



The endless human–machine monitoring loop of human bodies, minds and hearts in driverless cars

According to the National Transportation Safety Board's investigation of the Uber accident that resulted in the death of Elaine Herzberg in Tempe, Arizona, in March 2018, the test driver in the Uber spent 34 percent of her time looking at her phone streaming the TV show *The Voice*.[1] In the three minutes before the crash, she glanced at her phone 23 times. And, "The operator redirected her gaze to the road ahead about one second before impact." This is known because the Volvo Uber she was test-driving was fitted with a driver-facing camera. In fact, driverless cars today are fitted with a variety of cameras, sensors and audio recording equipment to monitor the human drivers of almost-driverless cars. Why should human drivers be under surveillance? And if a car is driverless, why does it need a human driver?

The autonomous vehicle (AV) is supposed to drive itself. This means it can navigate a path between two points and make decisions about how to deal with things that happen on that path. However, there is still some distance to go before this is technically feasible. Currently, semi-AVs require human drivers to be alert and vigilant, ready to take over at a moment's notice should something go wrong. That's exactly what the Uber test driver did not do. Nor did the three other test drivers in three fatal accidents involving semi-AVs. In each of the accidents, the driver did not take back control from the semi-AV because he or she was distracted by something else; ironically, the driverless car is supposed to free up the human to do other things. I refer to this as an *'irony of autonomy,'* playing on what researcher Lisanne Bainbridge wrote about automation in 1983: "The automatic control system has been put in because it can do the job better than the operator, but yet, the operator is being asked to monitor that it is working effectively."[2]

And now the loop of the human-and-machine has become a spiral. The human manager who oversees the semi-AV is overseen by a different kind of technology: affective computing. Affective computing is an applied and interdisciplinary field that analyzes individual human facial expressions, gait and stance to map out emotional states. By associating every single point on the face and how it moves and looks when conveying a particular emotion and combining the findings with posture and gait, affective computing can allegedly tell what a human is feeling. However, after a review of a thousand studies, psychologists brought together by the American Psychological Association found "Efforts to simply 'read out' people's internal states from an analysis of their facial movements alone, without considering various aspects of context, are at best incomplete and at worst entirely lack validity, no matter how sophisticated the computational algorithms."[3]

Despite this, the "emotional AI" industry is estimated to be worth $20 billion. Affectiva, an emotional-measurement technology company, writes that its product "understand[s] drivers' and passengers' states and moods … to address critical safety concerns and deliver enhanced in-cabin experiences … unobtrusive measures, in real time, complex and nuanced emotional and cognitive states from face and voice." Affective computing can be used to understand drivers' and passengers' states and moods as well as issues like road rage and driver fatigue. During the past year, I have identified 47 affect patents—patents for innovations that register the affective and psychological states of humans in vehicular contexts. Patents can signal intent to markets, customers and competitors rather than conclusively verify the state of a technology. Still, they make for fascinating reading. Patent pending number JP-2005143896-A, for instance, proposes to "determine the psychological state of a driver." Its telematics sensors create data from (1) the force with which a driver steps on the brake and (2) the torque applied to the steering wheel when turned. Patent number US-2017294060-A1 uses an on-board diagnostic system to record the driver's behavior and give real-time advice about how to drive in a fuel-efficient manner.

affective computing can be used to understand drivers' and passengers' states and moods as well as issues like road rage and driver fatigue

The monitoring of human drivers in semi-AVs is likely to increase both for reasons of safety and for the management of future insurance and liability claims. Thus, the irony is that autonomy for machines is not in fact a real separation of human and machine, as it is often viewed, but is instead enabled by human bodies and minds and monitored and managed by computing programs—even as the humans maintain the fantasy of machine autonomy.

1. National Transportation Safety Board public meeting, November 19, 2019. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian; https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx
2. Lisanne Bainbridge. Ironies of Automation. *Automatica* 1983;6:775-779; https://www.ise.ncsu.edu/wp-content/uploads/2017/02/Bainbridge_1983_Automatica.pdf
3. LF Barrett, R Adolphs, S Marsella, AM Martinez, SD Pollak. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychol Sci Public Interest* 2019;1:1-68; https://doi.org/10.1177/1529100619832930

Mapping an AI future

Technological advances have always disrupted society. But AI brings a step change from prior technologies for how it influences the nature of our interactions, with significant if not yet fully foreseen or understood consequences for how we organize our societies. AI's influence is outsized for its augmentation of human capabilities, for its challenges to what it means to be human and for its creation of a set of human–machine interactions that are qualitatively different from those of the past.

for the first time, we are adjusting to a reality in which one party has no agenda of its own

AI in the abstract is merely a complex system built on trust between humans and machines. For the first time, we are adjusting to a reality in which one party has no agenda of its own. Forging these bonds of trust will require us to push the boundaries of rulemaking for a broader lens and a tighter but flexible grip over this new, complex system.

To begin to lay the foundations for mapping this collective effort, we had the privilege of convening the world's leading thinkers. They helped us identify emerging issues and key questions such as:

+ *How will we manage the accountability for decision-making systems that increasingly complement—or even replace—human judgment and address their ethical considerations when we may not understand the implications of their algorithms?*
+ *How might algorithmic decision-making impact inequality and inclusion?*
+ *In what ways will consciousness, memory and emotion, which are now jointly created and expressed by human and artificial systems, influence social and legal norms?*

Many of our social rituals and legal rules are based on commonly agreed definitions of personal rights, corporate law and who ultimately bears responsibility for individual and societal well-being. These enshrined relationships of power and justice that reflected the realities of their day are threatened today due to ongoing technological adaptation. Given the already apparent disruption of these fundamental ideas, the rapid pace of technological *and* societal change and the inability of our current rule-making systems to keep up, what are the mechanisms by which we can proactively shape a new set of rituals and rules to help us prepare for and wring more societal value from a technology-enhanced society?

With more questions than answers after our convening, much work remains. That is why later this year, we plan to bring together a group of thinkers who will begin mapping out the contours of a new AI governance system.

By framing the most important AI-linked issues we will face in the next 10 years, we can make commitments and procure resources to advance a new governance of AI that addresses this new paradigm.

These conversations serve as a starting point: the beginning of a dialogue—not only a dialogue among those currently engaged but one expanded to a broader audience. To reach that audience, we seek entry points and metaphors to create opportunities for those at both ends of the spectrum—those who believe AI will transform society for the better and those for the worse—to voice considerations that are most critical from their perspectives. Please join us on this journey.

Hunter Goldman – Director, Innovation,
The Rockefeller Foundation
@huntergoldman

Contributors and Residents of A Month of AI in Bellagio, 2019

amir baradaran

is an artificial intelligence (AI) artist, visiting scholar and arts-based researcher at Columbia University. He is the founder of iBEGOO—software that democratizes the creation of AI-enabled augmented-reality experiences. Amir is a TEDx speaker and recipient of awards from the Knight Foundation and the Canada Council for the Arts. He also founded An[0]ther {AI} in Art, a platform to decolonize the future of AI and art making.

@amir_baradaran

tim davies

is director of social justice-focused consultancy Practical Participation and an independent researcher and practitioner focused on the social and political impacts of data policy—particularly policy around open data. Previously, he led a global research network with the World Wide Web Foundation on open data in developing countries, and designed and led its Open Data Barometer project, which tracks the pervasiveness and impact of global open-data initiatives.

@timdavies

maya indira ganesh

is a technology researcher and writer with a hybrid portfolio of work across cultural organizations, academia and nongovernmental organizations (NGOs). Her doctoral research at Leuphana University of Lüneburg in Germany investigates the interaction of computational techniques with cultural narratives in the construction of machine autonomy; AI; and the evolving role of the human in those areas. Maya comes to academic research after more than a decade with nonprofits.

@mayameme

claudia juech

is founding CEO and a board member of the Cloudera Foundation in Silicon Valley. Working with partners around the globe, Claudia and her team identify and support large-scale, data-driven opportunities to address the world's pressing challenges in any area of critical need, including health, the environment, economic inequality and education. Previously, she was associate vice president at The Rockefeller Foundation, leading the organization's Strategic Insights Division.

@cjjuech

nils gilman

is vice president of programs at the Berggruen Institute, a research organization that develops ideas to shape political, economic and social institutions for the twenty-first century. There, he leads the institute's research program directs its resident fellowship program, and serves as deputy editor of its publication, the *WorldPost*. He was previously associate chancellor at UC-Berkeley, where he earned a BA, MA and PhD in history.

@nils_gilman

hilary mason

is Data Scientist in Residence at Accel Partners, former General Manager of Machine Learning at Cloudera, and Founder and CEO of Fast Forward Labs, which was acquired by Cloudera in 2017. She's on the board of the Anita Borg Institute and advises several startups. Mason served on Mayor Bloomberg's Technology Advisory Board, and is a member of Brooklyn hacker collective NYC Resistor.

@hmason

Continued on next page

contributors

sarah newman

is senior researcher and principal at metaLAB (at) Harvard and Berkman Klein Center for Internet & Society fellow. As a researcher expressing ideas through installation art, she engages her work with technology's role in human experience. In addition to her art practice, she is a facilitator and educator and leads workshops that use creative materials to address interdisciplinary research problems. Sarah Newman is a 2017 AI Grant fellow and a cofounder of the Data Nutrition Project.

@sarahwnewman

tim o'reilly

is founder, CEO and chairman of O'Reilly Media, a company that has been providing the picks and shovels of learning for the Silicon Valley gold rush for the past 35 years. He is also a partner at early-stage venture firm O'Reilly AlphaTech Ventures and is on the boards of Code for America, *PeerJ*, Civis Analytics, and POPVOX. He is the author of *WTF? What's the Future and Why It's Up to Us*. Tim is working on a new book about why we need to rethink antitrust in the era of Internet-scale platforms.

@timoreilly

jake porway

is CEO of DataKind, a nonprofit dedicated to harnessing data science and AI to serve humanity. DataKind teams more than 20,000 global pro bono volunteers and experts from academia and industry with visionary changemakers to collaboratively design innovative data and AI solutions. They have tracked illegal mining from satellite imagery, delivered water more efficiently to drought-stricken regions with predictive analytics, and helped increase community college graduation rates.

@jakeporway

marietje schaake

is international policy director at Stanford University's Cyber Policy Center, international policy fellow at Stanford's Institute for Human-Centered AI, and president of the Cyber Peace Institute. From 2009 to 2019, she served as a member of European Parliament for the Dutch Liberal Democratic Party, where she focused on trade, foreign affairs and technology policies. She also writes a monthly column for the *Financial Times* and a bi-weekly column for *NRC Handelsblad*.

@marietjeschaake

katarzyna szymielewicz

is a lawyer specializing in human rights and technology. She is president of the Panoptykon Foundation, a Polish NGO that defends human rights in a surveillance society, and vice president of European Digital Rights. Her engagement in politics leading to GDPR's adoption was featured in David Bernet's 2016 documentary *Democracy*. She holds degrees in law and development economics from the University of Warsaw and London University, and is a 2015 Ashoka Fellow.

@szymielewicz

stefaan verhulst

is cofounder and chief of R&D at NYU's Tandon School of Engineering's GovLab, where he builds evidence on how to transform governance through data and collective intelligence. Stefaan is also editor in chief of *Data & Policy* and curator in chief of the Council of Europe's Living Library, an online curation platform. Earlier, he cofounded Oxford University's Comparative Media Law and Policy Programme and served as the UK's UNESCO chairholder in communications law and policy.

@sverhulst

richard whitt

is a corporate strategist and technology policy attorney. He is currently fellow in residence with the Mozilla Foundation and senior fellow with Georgetown Law's Institute for Technology Law & Policy. As head of consultancy NetsEdge LLC, Richard advises companies on complex governance challenges at the intersection of market, technology and policy systems. He is also president of the GLIA Foundation and founder of the GLIAnet Project. He is an 11-year Google veteran.

@richardswhitt

andrew zolli

oversees global impact initiatives at Planet, a company that provides satellite imagery data for real-time insights on business and social challenges. He ensures that Planet's data, products and services achieve their highest humanitarian, sustainable development and scientific potential. Andrew works closely across the company and with the UN, NGOs, philanthropies, conservation organizations and disaster-response entities to incubate and deliver new forms of social and environmental impact.

@andrew_zolli

about

The Rockefeller Foundation

advances new frontiers of science, data, policy and innovation to solve global challenges related to health, food, power and equity and economic opportunity. As a science-driven philanthropy focused on building collaborative relationships with partners and grantees, The Rockefeller Foundation seeks to inspire and foster large-scale human impact that promotes the well-being of humanity throughout the world by identifying and accelerating breakthrough solutions, ideas and conversations. For more information, visit rockefellerfoundation.org.

The Bellagio Center

is a hub for innovation, expansive thinking and cross-disciplinary practice that expands The Rockefeller Foundation's capacity to catalyze and scale transformative ideas, create unlikely partnerships that span sectors, and take risks others cannot. For over six decades, the Center has convened prominent experts, influencers, and other key stakeholders to spread knowledge, form new partnerships and financial commitments, and advance initiatives that support the Foundation's goals. Through its residency and conference programs, the Center has a long legacy of stimulating critical dialogue, thinking and actions that have made major contributions to the Foundation's enduring mission, "to promote the well-being of humanity throughout the world."

cre+dits

Dor Glick
The Rockefeller Foundation
Project Manager

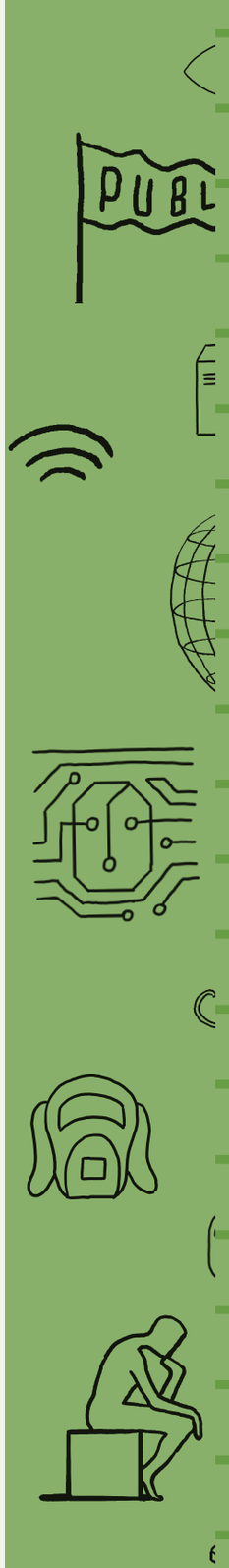
Meredith Kellner
Special Advisor

Carolyn Whelan
Editor

50,000feet
Design

Nicholas Blechman
Illustrations

Views expressed in AI+1 are those of the contributors, not necessarily The Rockefeller Foundation



AN INDUCTIVE INFERENCE MACHINE

With this meager apparatus, it becomes possible to start work immediately on what is believed to be the heart of the artificial intelligence problem: i.e., how to get new good ideas from old ones.

R.J. Solomonoff, participant,
Dartmouth Summer Research Project (1956)


The
ROCKEFELLER
FOUNDATION