CUTTER Business Technology Journal

Management, Innovation, Transformation

Vol. 32, No. 8, 2019

CAUTION!

AI Consequences Ahead

Lou Mazzucchelli
Guest Editor

CUTTER Business Technology Journal

As business models for creating value continue to shift, new business strategies are constantly emerging and digital innovation has become an ongoing imperative. *Cutter Business Technology Journal* delivers a comprehensive treatment of these strategies to help your organization address and capitalize on the opportunities of this digital age.

Cutter Business Technology Journal is unlike academic journals. Each monthly issue, led by an expert Guest Editor, includes five to seven substantial articles, case studies, research findings, and/or experience-based opinion pieces that provide innovative ideas and solutions to the challenges business technology professionals face right now — and prepares them for those they might face tomorrow. *Cutter Business Technology Journal* doesn't water down or delay its content with lengthy peer reviews. Written by internationally known thought leaders, academics, and practitioners — you can be certain you're getting the uncensored perspectives of global experts.

You'll benefit from strategic insight on how the latest movements in digital innovation and transformation, artificial intelligence/machine learning, Internet of Things, blockchain, analytics, and cloud, to name a few, are changing the business landscape for both new and established organizations and how cutting-edge approaches in technology leadership, enterprise agility, software engineering, and business architecture can help your organization optimize its performance and transition to these new business models.

As a subscriber, you'll also receive the *Cutter Business Technology Advisor* — a weekly bulletin featuring industry updates delivered straight to your inbox. Armed with expert insight, data, and advice, you'll be able to leverage the latest business management thinking to achieve your organization's goals.

No other journal brings together so many thought leaders or lets them speak so bluntly — bringing you frank, honest accounts of what works, what doesn't, and why. Subscribers have even referred to *Cutter Business Technology Journal* as a consultancy in print and likened each month's issue to the impassioned discussions they participate in at the end of a day at a conference!

Get the best in thought leadership and keep pace with the technologies and business models that will give you a competitive edge — subscribe to *Cutter Business Technology Journal* today!

Founding Editor: Ed Yourdon
Publisher: Karen Fine Coburn
Group Publisher: Christine Generali
Managing Editor: Cindy Swain
Copy Editors: Jennifer Flaxman, Tara Meads
Production Editor: Linda Dias
Client Services: service@cutter.com

***Cutter Business Technology Journal®** is published monthly by Cutter Information LLC, 37 Broadway, Suite 1, Arlington, MA 02474-5552, USA • Tel: +1 781 648 8700 • Fax: +1 781 648 8707 • Email: cbtjeditorial@cutter.com • Website: www.cutter.com • Twitter: @cuttertweets • Facebook: Cutter Consortium. ISSN: 2475-3718 (print); 2475-3742 (online).*

Subscription rates are US $485 a year in North America, US $585 elsewhere, payable to Cutter Information LLC. Reprints, bulk purchases, past issues, and multiple subscription and site license rates are available on request.

NOT FOR DISTRIBUTION
For authorized use, contact Cutter Consortium +1 781 648 8700 or service@cutter.com.

☐ Start my print subscription to *Cutter Business Technology Journal* ($485/year; US $585 outside North America).

Name _____ Title _____

Company Address _____

City _____ State/Province _____ ZIP/Postal Code _____

Email (Be sure to include for weekly *Cutter Business Technology Advisor*) _____

Fax to +1 781 648 8707, call +1 781 648 8700, or send email to service@cutter.com.
Mail to Cutter Consortium, 37 Broadway, Suite 1, Arlington, MA 02474-5552, USA.

Request Online License Subscription Rates

For subscription rates for online licenses, email or call: **sales@cutter.com** or **+1 781 648 8700.**

CUTTER CONSORTIUM
Access to the Experts

Opening Statement

by Lou Mazzucchelli

AI: The Third Time Is Not the Charm …

Let's begin with some disclosure: I have always been fascinated with the concept of artificial intelligence (AI). As a college undergraduate, I immersed myself in the fundamental technologies that defined the field at the time and designed an independent study program that resulted in the first bachelor's degree granted in AI in 1977.

I did not aspire to academia, however, and one of my better career moves was *not* pursuing AI in industry after graduation. Nonetheless, I did keep up with the technology over the years, wondering if the possibilities explored at school might ever be instantiated. Since college, I have experienced three AI "waves." Let's take a closer look at them before delving into the thoughtful articles of this issue of *Cutter Business Technology Journal (CBTJ)*.

Wave 1: LISP

The first wave, in the late 1970s, was driven by the emergence of LISP machines. For the uninitiated, LISP was the AI programming language of choice back in the day (and remains a personal guilty pleasure), but its runtime performance was notoriously slow, partially because its language grasp exceeded the hardware reach of general-purpose computer implementations. Lisp Machines, Inc., and, later, Symbolics both attempted to create markets for specialized hardware that would run LISP code faster than general-purpose hardware and, therefore, enable the creation of "real" AI applications and systems.

I remember talking with Patrick Winston at Symbolics, who blithely remarked that, "in the future, every computer will be a LISP machine." (He was right, but only because general-purpose architectures later got fast enough to run LISP at scale.) One of my industry scars (or, perhaps, a badge of courage) is when Cadre Technologies, a global software design tools/environment supplier company that I founded back in 1982, purchased a Symbolics computer, at my insistence, to facilitate a joint project with General Electric (GE) to commercialize a LISP-based AI system for Ada software development. As with many projects of the period, this one was ambitious but met with limited success (the only road to commercialization of this project was to reimplement it in another, more portable, language).

In the wake of the visible and costly failures of LISP machines, the first AI wave subsided. This does not mean that AI disappeared; rather, it merely went offstage for a bit while people worked on it in the wings.

> *Today, machine learning (ML) leads the way for the third AI wave.*

Wave 2: Natural Language

Natural language recognition ushered in the second AI wave. Drastically increased value in terms of CPU performance per dollar allowed designers to create devices that would do a reasonable job of dealing with voice in ever-less-restricted domain areas. With further enhancement of this technology, it became the harbinger of the third, and current, AI wave.

Wave 3: Machine Learning

Today, machine learning (ML) leads the way for the third AI wave. ML systems generate responses using directed pattern-matching and feedback to satisfy a goal, like "detect an edge," or "signify whether this is a picture of a cat," or "indicate whether this human is a felon." These systems gain "skill" by ingesting ever-larger data sets ("training"). And we feed the systems

ever-larger data sets to improve their training. We have observed some spectacular results (e.g., world-class *Go*-playing systems "grown" in weeks). However, we are at a loss to explain, at a micro level, exactly how these systems make decisions.

> *We are placing a lot of trust in ML systems that are increasingly running the joint, from Google pet searches to hiring decisions.*

In the human world, accepting a decision without questioning the facts leading up to it is a definition of "trust." By that definition, we are placing a lot of trust in ML systems that are increasingly running the joint, from Google pet searches to hiring decisions. This brings several problems into our view, leading to some important questions:

1. Where is the record of initial training approaches for any ML system? We cannot be sure that bias has not been introduced into the system without knowing the initial conditions and weights of the data sets.
2. Where is the record of changes in response to training input, and how is that input supplied or collected? Who gets to decide?
3. Where does liability lie for ML mischaracterizations?

While we can laugh at stories about ML misidentifying members of the US Congress as criminals (an easy mistake?),[1] life becomes more difficult for someone denied a job or promotion by an opaque ML-based system.[2]

These thoughts, among others, prompted my interest in an issue of *CBTJ* that would explore the state of AI from a mostly nontechnical perspective, focusing on emerging ethical challenges that we will face, or ignore. While the issue was being prepared, I've been noticing others raising similar questions. For example, Rich Caruana and his team at Microsoft Research are focusing on "intelligible, interpretable, and transparent"[3] machine learning, something our first author, Cutter Consortium Fellow Lynne Ellyn, also points out. Moreover, leading data analytics firm SAS has recently released a white paper, "Machine Learning Model Governance,"[4] that describes a process to manage some of the issues we explore in this issue. However, while the process described in the white paper asks "what, when, and how" ML models are used, it omits the question of "why?"

Upcoming Topics

Digital Architecture
Gar Mac Críosta

Fintech/Blockchain
Karolina Marzantowicz

In This Issue

The contributions in this issue of *CBTJ* will help us get up to speed with the current state of AI and to think about some of the issues raised when we look beyond systems that appear to work as intended. Our contributors span industry and academia, and their commentary provides a good way to gain an overview of the problem.

We begin the issue with an article by Lynne Ellyn in which she recounts her experiences with AI technology in the real world, surveys the current landscape, and identifies key nontechnical issues that companies are likely to face when deploying AI-based systems.

From Ellen's on-the-ground view, we then go to outer space (well, low Earth orbit, actually) to examine the issues around AI (in its ML incarnation) employed in a NASA system to track orbital debris. In his article, William Jolitz, the inventor of OpenBSD (open source Berkeley Software Distribution), makes the case for organization-wide awareness and alignment around ML and suggests that, like security, transparency cannot be bolted on later; it must be addressed at a project's origin.

Experienced IT practitioners know that errors will occur. A big part of building and managing complex systems is dealing with risk management (which includes identification and mitigation strategies). This is hard enough when documentation and source code exist. But the current state of ML-based AI tends to result in opaque black boxes, which make this activity, um, challenging. This brings us to our next article by David Biros, Madhav Sharma, and Jacob Biros, who

explore the implications for organizations and their processes.

One way of getting an off-course system (or person) back on track is by nudging. This concept can be particularly useful in goal-directed systems. But, to reiterate, errors will occur. In his article, Richard Veryard describes technologically mediated nudging; the possible unintended consequences; and the need to consider the planning, design and testing, and operation of the system for robust and responsible nudging.

As AI becomes more visible as a corporate strategic tool, organizations will have to incorporate issues surrounding AI as part of corporate strategy. Pavankumar Mulgund and Sam Marrazzo help us by providing a framework for developing an AI strategy. The authors discuss the "minimum viable model" approach to the development of the underlying AI/ML models, along with the platform on which these models run and the inevitable tradeoffs. They conclude their piece by examining some best practices for the successful implementation of AI initiatives.

In the closing article, Cutter Consortium Senior Consultant Paul Clermont describes some of the impact that AI has had at the boundaries of commercial organizations and public policy in an article aptly entitled, "Who Knew THAT Would Happen?" Those of us who have experienced unintended consequences of other technologies will want to answer "anybody" but should remind ourselves that some may not have the memory of prior years, and that hindsight is perfect. Clermont explores how to identify possible unintended consequences in advance and proposes countermeasures to negative unintended consequences in the form of design principles and public policies.

It's my hope that this collection of papers can help you "set the table" for further exploration, discussion, and policy development in your organizations. I firmly believe that creators and implementers will either get ahead of these issues, or regulators will act after the fact. Or, perhaps, the free-market rush to AI market dominance will lead our society implicitly to the same place that communist China is heading explicitly. Is that the future we want?

Perhaps the most dangerous aspect of the third wave of AI is its apparent efficacy. As *Young Frankenstein* taught us, putting a monster in a tuxedo may make it appear less threatening but likely only masks other problems.

References

[1]Bayern, Macy. "Amazon AI Misidentifies Congress as Criminals, Proves It's Not Ready for Enterprise." TechRepublic, 27 July 2018.

[2]Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women." Reuters, 9 October 2018.

[3]Caruana, Rich, et al. "Intelligible, Interpretable, and Transparent Machine Learning." Microsoft, 2019.

[4]Asermely, David. "Machine Learning Model Governance." SAS Institute, 2019.

Lou Mazzucchelli is a Fellow of Cutter Consortium's Business Technology & Digital Transformation Strategies and Data Analytics & Digital Technologies practices. He provides advisory services to technology and media companies. Early in his career, Mr. Mazzucchelli spent 13 years leading Cadre Technologies, a pioneering CASE tools company that he founded in 1982, which grew to become one of the top 50 US ISVs before its sale in 1996. During this period, he was listed in the "Top 200 in the Software Industry" by Software Magazine *and was a member of the Airlie Council, a group of US software engineering experts and thought leaders impaneled by the US Congress to drive reform in software development and acquisition practices.*

Since his work at Cadre, Mr. Mazzucchelli has held various positions, including Venture Partner at Ridgewood Capital, where he helped build and manage its technology portfolio; interim CEO at LightSpace Technologies, a pioneering 3D visualization company; and Director of Asure Software, a US public company. Prior to these roles, he served as a technology investment banker and equity research analyst at Gerard Klauer Mattison. Mr. Mazzucchelli was once named to the Wall Street Journal *all-star team and was one of nine "Home-Run Hitters" analysts (out of 2,400) recognized for his stock-picking performance. He began his career in data communications, moving to IT management and consulting before founding Cadre. He can be reached at lmazzucchelli@cutter.com.*

A MACHINE IS STILL A MACHINE

The AI Journey: What Is Real, and What Is AI?

by Lynne Ellyn

Everywhere these days, we are bombarded with ads from companies telling us what amazing results they are producing with artificial intelligence (AI) and the Internet of Things (IoT). These ads even pop up on my iPhone while playing *Wordscapes* or *Words with Friends*. And I recently heard a feature on NPR (National Public Radio) about two products *invented* by AI — software inventing tangible products (something I have yet to verify). So the question we must now ask ourselves is, "What is real, and what is AI?"

In the early days of computing, we were enthusiastic about the ability of computers to automate routine but labor-intensive tasks like calculating payroll for thousands of employees or managing airline scheduling. Remember Sabre? Many of you reading this will not remember but, trust me, in its day Sabre was a marvel of technical advancement. American Airlines, where Sabre was developed, had every other airline scrambling to duplicate its ability to schedule flights and maintenance and maximize load (profit).

Sabre was an innovative algorithm that managed a sophisticated and complex process that previously required the attention of hundreds of people. While not labeled AI, Sabre turned the travel industry upside down. It was another step in the progression of greater sophistication in computing algorithms — a program making decisions that previously required humans.

So is AI merely more sophisticated algorithms, or is it something more? Is AI a new technology, or the evolution of an existing technology? Is the Turing test still the way to determine whether a software program (or a robot) is intelligent? IBM's Deep Blue can beat chess masters using massive processing capability. It might meet the Turing test, but is it intelligent?

From Then … Till Now

Back in the late 1980s and early 1990s, I was leading a group at Chrysler Corporation focused on advanced technology and AI. We successfully deployed a natural language AI product for database access, a forward chaining inference product for vehicle configuration management, a neural network product for detecting calibration "drift" or errors in machines used in the assembly of trucks, a neural network to detect warranty fraud, and a host of other AI projects (more than 150 deployments). We used backward and forward chaining systems, fuzzy logic, neural networks, and pattern matching. These are the methods that defined AI in the 1980s and 1990s. (Some of them actually date back to the late 1950s!)

In the early 1990s, I went to work at Xerox, which had a rich history of doing AI projects and research. While I was managing the software development group, we deployed a sales territory configurator based on a forward chaining inference engine, a system utilizing picture recognition using neural networks and an inference engine, and a neural network and fuzzy logic system that helped configure printer interfaces to software.

Fast-forward to 2019. What is the new *new* thing in AI? What new methods or approaches have been invented? Based on my research, it appears that the real news in AI is *processing*: advances in computer speed, more solid-state technology, small embedded neural networks, and multilayer neural networks. So what does all of this mean?

A Step Toward Deep AI

Neural network technology — aka "deep learning," a subset of machine learning (ML) — has advanced artificial intelligence to multilayered input and output connection layers. Older architecture limited the depth of learning that was possible. In fact, Marvin Minsky (the father of robotics and an early AI researcher) and Seymour Papert documented this limitation in their 1969 book *Perceptrons*.[1] Rather than advancing the field of AI, however, as it should have, the book's publication led to what was labeled the "AI winter" because funding for neural network research dried up as people misinterpreted Minsky's and Papert's comments to mean that neural networks didn't work. The limited neural networks of the time needed layers of perceptrons, which Minsky and Papert understood. In more

recent times, R&D has provided neural network software with multiple layers of connections. This software is not yet on the scale of a human brain (the brain has billions of connections) but is much more capable than the AI of the past.

With the development of these broadened and enhanced ML capabilities, we are now able to tackle more complex problems and more voluminous data. Indeed, people are now speaking of *deep AI* (when the neural network learns without supervision by a person). Is this new? Not really. The warranty fraud system that my team developed at Chrysler years ago was an early example of deep AI, or unsupervised learning. Basically, a neural network runs through complex data records thousands of times, clustering records that are similar in much the same way as do nearest neighbor data clusters and multivariate cluster analysis. With the increased sophistication of today's neural network technology and the massive increases in computing power, however, we can now use deep AI to tackle previously intractable problems by finding patterns hidden in a massive amount of data.

Let's return for a moment to the Chrysler warranty fraud system, where analysts inspected the outlier clusters and auditors were then sent into the field to examine the dealership records and interview the vehicle owners. Voilà — most of the outlier clusters contained fraudulent claims. The problem of warranty fraud was a difficult problem at the time, and the deployment of the neural network helped the corporation detect fraud and recover funds. The neural network, however, didn't know that the outlier clusters it had identified contained fraud (the neural network didn't actually *know* anything). The neural network at that time was only able to process a massive amount of data and find hidden anomalies, anomalies that were significant and subsequently proven to contain a disproportionate number of fraudulent warranty claims. But more importantly, the neural network highlighted what changes should be put in place so that the warranty claims system could prevent fraud.

The idea behind neural networks dates back to the 1960s. Neural networks simulate our understanding (then and now) of how the human brain learns. In the brain, the connections between synapses is a chemical/electrical activity. The more a neural connection fires off in the brain, the thicker the connections become. In software, this is simulated by the strength of the connection, which is, basically, a number. The neural networks of today are certainly more sophisticated and capable than those of the past, but they are, in essence, bigger and faster implementations of well-worn concepts.

Despite the amazing capability of neural network technology, we must remember that neural network software is software — an algorithm written by software developers — a program that "learns," whether supervised or unsupervised. It is also important to note that "learning" in the ML context is a change in system response based on statistical characteristics of the input data provided. There is no aha moment inside an ML system. Insight remains beyond our algorithmic reach.

People are now speaking of deep AI (when the neural network learns without supervision by a person). Is this new? Not really.

Rules-based inference engines have also been around for a long time. There are two basic types of inference engines, both of which are useful: (1) forward chaining inference engines, which are data-driven, and (2) backward chaining inference engines, which are goal-driven. Forward chaining is used for planning- and configuration-type tasks. Backward chaining is used for diagnostic or prescriptive systems. We can employ forward chaining to plan routes, configure vehicle options for profitability, determine store locations, and so on. Backward chaining is useful for problem analysis such as medical diagnosis or equipment failures. Rules engines that provide both methods can handle much more difficult problems than can one method alone.

Combining various AI techniques with massive data processing on very powerful computers is the real advancement in the AI field. The capabilities of each AI technique are geared to a particular type of problem, but in real life there are many problems that require multiple approaches — that is where the big advances in AI are occurring. Coupled with the ability to rapidly process a staggering amount of data, today's AI systems can help tackle problems that were insurmountable just a few years ago.

Dangers of AI

While the capabilities of AI are incredibly useful and offer business and government the opportunity to solve knotty problems, a common question is, "Does

this capability pose threats to humans?" Are the sci-fi scenarios of malevolent robots or killer software programs a real concern?

Let's focus on what we know about software. Software is the most complex product on earth. Unlike the complexity of, for example, a nuclear power plant, a bullet train transportation system, or a space shuttle, software cannot be seen (other than by reading its code), heard, touched, or measured effectively. The problem with inspecting code is that the brains of even extremely intelligent people can deal with only a finite set of items at once; in general, this limit is seven to nine items. A sophisticated software program will have millions of lines of code. No human can hold that many actions or conditions in his or her brain at any one time. While people can write a software program line by line, understanding the finished code, which may consist of thousands or millions of lines, is impossible.

> *Turning over critical systems to any kind of software is risky.*

Software is also subject to unforeseen *emergent* behavior (i.e., behavior that the developers did not anticipate; behavior that occurs when a circumstance that no one ever thought of presents itself to the software). Imagine the AI self-driving car of the future that has been programmed to deal with kids on bicycles, other vehicles, bad weather, road obstacles, and literally thousands of other scenarios. But what about scenarios that the developers never considered because they seemed implausible: a tsunami heading toward the road, for example, or an avalanche or a plane landing or crashing on the highway? Will that self-driving car just proceed because it has not been programmed to handle the scenario it now faces? Or will it react with programming that was aimed at other scenarios? Maybe it will stop moving … or speed up, or run in circles, or display some behavior never seen before. Every serious software engineer knows that no matter how carefully constructed or how exhaustively tested, *software will fail at some point*, and, unlike humans, software fails catastrophically in most cases. Emergent behavior is rarely helpful. A person driving a vehicle and who sees a plane approaching a roadway and about to crash will most likely change direction, speed up, leave the roadway, or take some other evasive action despite never having experienced such an event before. People have the ability to instantly take a new action when life throws an unforeseen problem at them. Whether that new action is indeed the best action is certainly debatable, but we cannot say (or assume) for certain that software has any such capability.

The recent disasters Boeing experienced with its 737 MAX appear to be a classic emergent behavior problem.[2] The software that serves as the autopilot was supposed to grant control to the pilot in an emergency. According to reports about the crashes, the pilots could not gain control of the aircraft because the software did not relinquish control. Were the software developers rushed? Was there inadequate testing? Were faulty sensors to blame? I will leave it to the people working on the problem to figure out how the planes malfunctioned, but my point is this: Boeing did not intend to produce planes that would crash as they did, and software was at the heart of the problem.

Turning over critical systems to any kind of software is risky. A system based on neural networks is really tricky because no one truly knows all the scenarios that the software has "learned." While the developers can indeed inspect the program code, the actual neural network is much less transparent. For this reason, nuclear power plants have avoided software-managed operation, instead using software to *assist* humans in operation, with the ability for humans to operate manually. Weapons systems have also historically had human oversight. Any scenario that allows an artificially intelligent system to determine when to fire a missile is a dangerous scenario!

The US stock market "flash crashes" in 2010 and 2015 further illustrate the danger of allowing software to "make decisions." When multiple systems are watching trading activity on the stock market and each can automatically sell when a stock experiences a price change, a software equivalent of the harmonic resonance[3] problem can occur. A small change triggers a sale by one brokerage, that sale triggers sales in other brokerages, and soon all the systems are acting and reacting in ways that do not reflect the value of the stocks. In the 2010 crash, the stock market plummeted close to 1,000 points over the span of five minutes before rebounding back (but still down 3.2%) by the closing bell, with the flurry attributed to a possible software glitch.[4] Did this "market reset" simply mean "Never mind, we are declaring that the events did not happen"?

Now, if we were to have a software harmonic resonance problem with weapons systems from different countries, instead of with stocks, recovery would be much more difficult; perhaps even impossible. As we expand the use of AI, we must be mindful of just these possibilities. Software should never control life-critical systems. Systems that provide advice or identify potential problems can be useful in medicine or in assessing terrorist threats, but such systems should not be allowed to take action without human involvement when human life is at stake. Software taking independent action in these types of situations could be fully as threatening as those scenarios of malevolent robots or killer software programs, especially when we consider bias and lack of transparency.

The entrenchment of bias is yet another danger that AI, especially neural networks, presents. Confirmation bias in the extreme is happening as we speak on social media. Click on an article about or a picture of kittens, for example, and your Facebook page starts to show more kittens. Harmless, perhaps, but click on "news reports" of terrorist activity or crimes committed by minorities, and your feed will be crammed with more of the same. This use of AI (for it is AI that determines what to show you in these cases) poses a danger to society by reinforcing biases and isolating people from fair and balanced news coverage and real knowledge. People become more extreme and entrenched in a particular view because the AI system learns what attracts someone's specific attention and then provides endless examples of that idea or point of view, ultimately providing that person only a very narrow window on reality.

The nature of neural networks obfuscates how and what has been "learned" by the software. There is a real need to improve the tools and the process so there is transparency as to how the results are derived. An example of this issue can be seen in a neural network that was very accurately determining whether a given picture depicted a dog or a wolf.[5] However, after much examination, it was determined that the neural network identified all pictures with snow as showing wolves but was not able to distinguish a wolf from a dog in photos where there was no snow. The training set had a hidden bias for linking wolves with snow, and snow became the identifying feature that the neural network learned.

The potential for neural networks to be discriminatory or have other negative bias is very high. Currently, there is no way to inspect a large neural network to determine how it is arriving at its conclusions. If the bias in the training data is matched by bias in the testing data, the results can look good but be very wrong. If neural network results are being applied to healthcare decisions or other decisions of great importance, it is therefore critical to explore the issue of bias. Until the technology becomes more transparent, companies deploying neural networks may find themselves in hot water with regulators and government officials since the decision process takes place in a "black box" without transparency.

The entrenchment of bias is yet another danger that AI, especially neural networks, presents.

Microsoft recently published a series of articles entitled "Intelligible, Interpretable, and Transparent Machine Learning,"[6] good articles stressing the inadequacies of different ML methods and ways to compensate for the inadequacies. But we need to add accountability to intelligibility and transparency. As you sally forth, excited about the possibilities, think intelligibility, transparency, and *accountability*.

Conclusion

So is AI a danger to society or a technical marvel propelling us into a future of programs that rival human intelligence? At its basic level, AI is about algorithms that handle data and produce results developed from models of how humans analyze and reason. AI systems have a great advantage in their ability to use these reasoning models at a scale that is orders-of-magnitude larger than that of a single person or even a group of people. As increasingly difficult problems are managed by AI, it will usurp tasks previously performed by highly trained people and become more "intelligent." But artificial intelligence is not (yet) at the same level as human intelligence. At present, AI can't do more than what a programmer or a team of programmers can envision, but it can do the imaginable at unimaginable speed, utilizing a massive amount of data. The results can be amazing and incredibly useful, but they could also be destructive if designed without ethical analysis and deployed without adequate oversight.

References

[1]Minsky, Marvin, and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 1969.

[2]"Boeing 737 MAX groundings." Wikipedia.

[3]"Harmonic Resonance." Audible-Acupuncture: Philosophy.

[4]Lauricella, Tom. "Market Plunge Baffles Wall Street." *The Wall Street Journal*, 7 May 2010.

[5]Bradbury, Danny. "You Should Find Out What's Going On in That Neural Network. Y'know They're Cheating Now?" *The Register*, 1 June 2018.

[6]Caruana, Rich, et al. "Intelligible, Interpretable, and Transparent Machine Learning." Microsoft, 2019.

Lynne Ellyn is a Fellow of the Cutter Business Technology Council and a Senior Consultant with Cutter Consortium's Business Technology & Digital Transformation Strategies practice. She retired in 2011 as the Senior VP and CIO at DTE Energy, a Detroit-based diversified energy company. During her 32 years in IT, Ms. Ellyn managed organizations with as many as 1,200 employees. She worked her way through the IT industry, beginning as a programmer, and has held such positions as systems analyst, project leader, knowledge engineer, and systems architect. Ms. Ellyn combines a strong hands-on technical background with an equally strong knowledge of all aspects of business. She has successfully held leadership positions in five industries: healthcare, automotive, high-tech, consulting, and energy. Ms. Ellyn completed a postgraduate certificate in the Foundations of NeuroLeadership with Middlesex University in Manchester, UK. She now focuses on leadership coaching, team development, IT strategy, public speaking, and consulting for Cutter Consortium. Ms. Ellyn also does equine-facilitated leadership training. She is passionate about building high-performance teams, coaching leaders for high performance, and managing complex business requirements. Ms. Ellyn has a bachelor's degree from Oakland University and an executive MBA from Michigan State University. She can be reached at lellyn@cutter.com.

IT'S ALL ABOUT THE STORY

Machine Learning and Business Processes: Transparency First

by William Jolitz

Machine learning (ML) isn't new, and enterprise computing certainly is accustomed to accepting the challenge of adopting new technologies and driving them hard to produce the best results. ML is opaque. Business processes are even more opaque.

What is new is that as ML is rapidly pushed into action across an organization, we cannot tolerate the risk that flawed technology introduction may injure brands and brand value or put a business's professionalism in question. Most earlier ML adoptions neither accepted nor needed thorough organizational transparency. They required only "point solutions" — a sharp, tight focus on a pain point to see how much benefit could be obtained. There was no reason to fear that a misapplication of ML might put an entire brand at risk, as the focus was on achieving the greatest benefit as applied to a specific problem. Point solutions have the advantage of aggressively adopting new technology of any stripe. But that "sharpness" can also have the unintended consequence of cutting an organization to ribbons by introducing a "pain" that wasn't there to begin with. Moreover, the point solution's lack of transparency hides the created pain. *Embedding transparency into how technology and business communities work together attacks this obstacle of ML misapplication.*

The best organizations extend transparency broadly throughout the organization (see sidebar, "Transparency Connects Cross-Functionally Alongside ML"). If you partition or compartmentalize transparency, for whatever reason (including "siloing"), you interfere with the hidden strengths that it brings. Among those strengths is resilience, arising from the quick discovery and correction of a flaw — because someone, somewhere, notices the flaw and has the answer to remedy it.

When employing new enterprise technology, it is a commonplace reaction to simply reproduce pilot projects, warts and all, to get results across the board. The intent is to replicate an initial success, but complex technologies, like rooted plants, don't always transplant well. Some bring along more or less than intended, introducing an undesired shock when the unspoken, prior presumptions change. This is an example of how, by accident, a powerful initiative to galvanize change laterally in an organization can instead lead to a loss in transparency. When one of those prior warts didn't scale, or the manner of scaling was at odds with the pilot, the wart wasn't really a wart at all, but rather a hidden, non-transparent, unaddressed obstacle that was allowed to pass.

The shock here is the pilot project's needful neglect of transparency as it wrestles with necessary expediency. We can avoid this shock by always retaining transparency at every step along the way — single-stepping through where stories, both prior and new, deviate. The benefit of maintaining thorough organizational transparency as a virtue is that we refine organizational engagement to capture unnoticed flaws as they briefly become visible.

Transparency Connects Cross-Functionally Alongside ML

Broad ML adoption is relatively new, and it impacts laterally across an organization. In addressing the accompanying transparency's lateral growth, it tends to bring along everyone else's "one more thing." Depending on the handling of such transparency, some may fear a loss of control. So how can organizations comfortably accomplish such broad ML adoption?

We can find a parallel in listening to music; sometimes adding another performer or instrument fills a void, or adds to clarity. But, in other cases, it just means there's more noise. Comfortable additions to ML adoption adjust the organizational "feedback" to keep it concise.

Of course, there's always a counterpoint to the harmony — in music or ML adoption. Many try to get a jump on the shifts they sense "in the wind" before they even happen. This calls for "scope" control.

Transparency is important to ensure the entire organization has the clarity and confidence of knowing where ML has helped, how it is helping, and where the boundary is between how the problem is being solved and the means the solution employs or doesn't employ to get results. *An organization-wide ML architecture integrated symmetrically across the organization, rather than integrated chaotically piecemeal, avoids intra-organizational conflict*. This organization-wide approach to unity is the power of a truly transformative approach — getting the "pull in" rather than the "push off."

Case Study: Making the Orbital Debris Elimination Complexity Process Transparent

Years ago, I built a minimal ML-generated model for a NASA Frontier Development Lab Challenge to apply machine learning to orbital debris elimination. I used open source examples from multiple repositories of different authors as a proof of concept. The model worked with supplied examples from NASA. However, when tested against broader cases beyond those supplied — unacceptable solutions that intersected the atmosphere and Earth's surface — it failed as a practicable model. Because of the compact model, however, NASA could be convinced that (1) it had found the correct solution, (2) the solution worked for the correct reasons, and (3) the proposed solution was incomplete as a final solution. There was immediate buy-in because the minimal model was transparent.

Figure 1 shows the complexity of the state-transition matrix representing the model's solution to accomplish

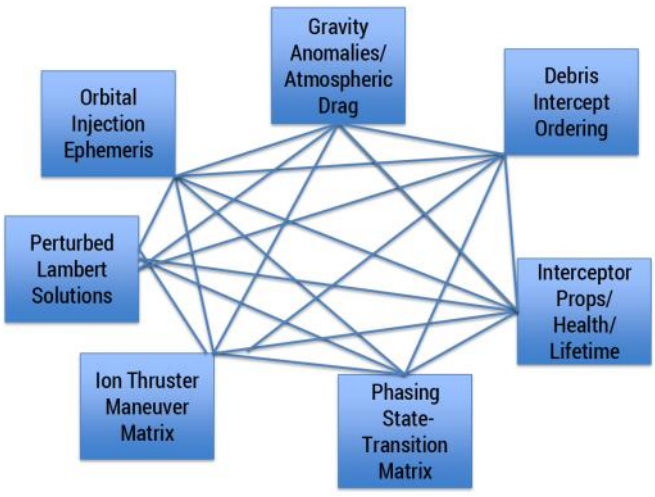

Figure 1 – State space orbital debris intercepts.

the debris elimination mission. The minimal model I designed correctly and completely answered this problem for the cases that were supplied.

It wasn't the machine learning itself, nor the problem definition, nor the resulting expanded model that was the concern in solving the challenge. The problem was how to convince both model implementers and stakeholders at every step along the way that the solution being incrementally proven (using engineering change management) from this, the most fundamental model possible, would end up as the valid one they sought.

The difficulty was ensuring transparency so that incremental changes in ML technique, as well as increasing constraints and complexity, were visible. Transparency allowed both implementers and stakeholders to see the impact of what they were doing as they built the rest of the solution.

So the consequential pilot project "warts" were due to creeping functionalism that added benefit obscurely. Transparency avoided the issue of difficulty in tracking when a numerical solver used an inadequately constrained result from another model. Many aerospace codes are carefully curated over a significant timespan for this reason. They are difficult to track and have large consequences. If the code were to be blindly copied by an elaborate means of any kind, we wouldn't be able to discern its vulnerabilities. So, no matter how fast and accurately its solutions might match an arbitrary number of test cases and actual cases, it might fail in an unfortunate way. *It was created non-transparently to begin with and without any knowledge that it would be used inside an ML framework that might change how it behaved.*

By starting with a "minimal" model for this challenge, transparency could be maintained from beginning to end. As such, both stakeholders and implementers could explain to policy makers the history of how this approach came to be, rather than have an obscure, large, "all up" implementation that could only be empirically tested *without explanation*.

By using aerospace's system engineering as a guideline, engineering change management (see Figure 2) controlled the successive iteration of the code so that when mistakes accidentally became incorporated, they could be backtracked when found and remedied.

NASA itself has had a history of requiring explanation (i.e., transparency). An example of this is illustrated in the movie *Hidden Figures*. In that retelling of the earliest

US crewed spaceflight, astronaut John Glenn trusted the calculations of the women working as human "computers" more than the actual computers' potential "garbage in, garbage out" calculations. In my own experience with NASA, countless hours were spent with a numerical simulation of flight navigation and guidance software before actual vehicle operation, where even round-off errors and the slightest errors in gyroscopes or trigonometric functions could accumulate to create consequences in use. People, using human decision-making skills, made the final decisions; policy makers, like astronauts, do not want to rely solely on machine algorithms or automated test suites. Ultimately, they want to be able to look a person in the eye when they ask, "Does it work?"

Stakeholders must be onboard long before the final decision because transparency is a level of trust that starts at the beginning and is carried along to the end. At the final stages of a program, you want the last thought to be, "Have we forgotten something?" rather than "Hoping on a wild-ass guess."

State of the Art

There's no common expectation of how to adopt machine learning across industry segments and verticals. Many sanguine experts warn of reasonable fears as a result. And there are better and worse ways to make the effort tractable, as well as the appropriate skills to address this as a data scientist.[1]

In various engineering applications that have long been dominated (and made vulnerable) by non-transparent mechanisms and processes, ML applications are handled just the same as all other opaque analytics done for decades without much oversight. *Business pays for this as a form of unconstrained liability* that bites unpredictably, leading to a collective shrug of, "What can be done?" as a form of lip service to transparency, and that the engineers will handle it "after the fact." This approach is irresponsible, as the first flight of the Space Shuttle STS-1 illustrates. The flight had many flaws, including wrongly locating the center of pressure by using an ideal gas model instead of a real gas model, and the two astronauts survived mostly by sheer luck.

At the other extreme are highly disciplined parts of the healthcare industry. An epitome of this is Stanford Medical Center's view on the subject. Stanford not only demands a detailed explanation, justified through existing and historical medical practice; it also wants any ML application's results proven on a case-by-case basis.[2] Stanford sees ML as too easily crossing ethical, legal, and medical "no go" boundaries that define the Center's practice and standard of care, leaving things "half-baked." Therefore, it wants ML just as "baked" as a fully justified and examined traditional medical practice. This dilemma motivated my interest in this topic, because medicine strives to "do no harm," including harm from the non-transparent, unproven parts of an ML "improvement."

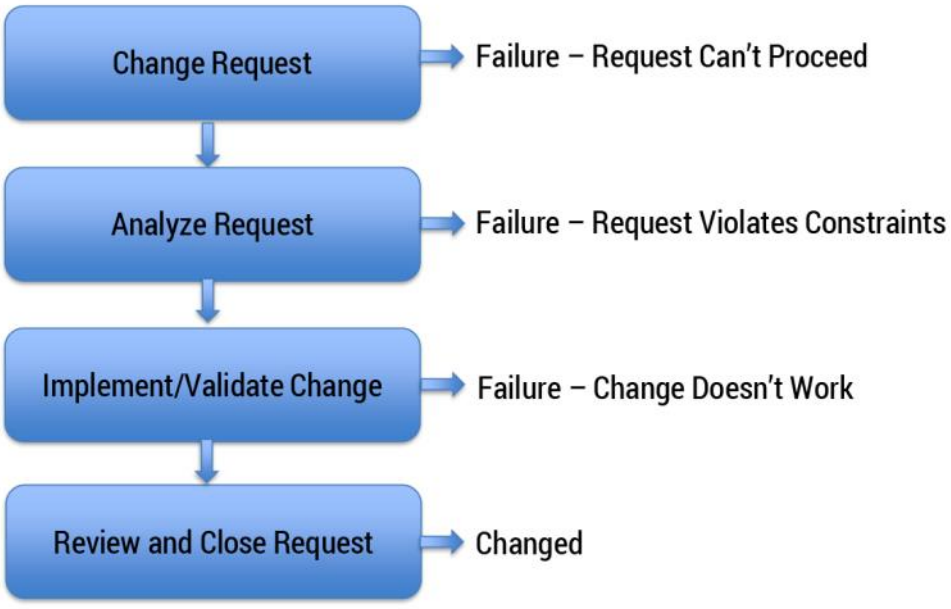

Figure 2 — Engineering change management.

The medical side is full of interesting examples for understanding best practices. Bias in an ML application can originate in human-created medical records,[3] revealing hidden conflicts between human specifications and ML algorithms. It seems errant to raise the flag on ML applications when they are replicating human-induced fractures to the system. Yet, we should also consider that all fractures are "human-induced"; after all, we're human, and we developed machine learning. And all fractures are, to paraphrase W. Edwards Deming, "working outside the system when it's broken."[4]

Our natural tendency is to compensate in opaque ways. However, through incremental, individually transparent steps, and with a change management process to handle any need to reverse or reconsider these steps in the future, these fractures, captured as shared (ML or medical or business practice) problem semantics and constraints, can be systematically addressed as they are encountered.

> *"Fit" and "feel" apply both to the problem and the means to solve it.*

This problem can be viewed more whimsically as "If the answer is '42', then what was the question?"[5] Having the answer runs counter to not knowing how to implement that answer. This can be likened to the organizational problem of invoking powerful new technology when you don't have a grasp of how to "let it loose" on your internal business, processes, and customers. The answer is to implement *carefully* and with *great transparency*, so that many in an organization "get" what, where, and how they are going. Only then can they decide whether that is the direction they want.

Building ML Models While Retaining Organizational Transparency

When building ML models, start with clear boundaries between the ML (and related technology/IT) community and the stakeholders they serve, and keep all participants knowledgeable of these boundaries throughout, beginning with statements of the problem(s) to be solved. The stakeholders must begin by knowing where they can and cannot go in solving a problem. The boundaries can be affected by subjective, objective, situational, transactional, and normative constraints, ranging from management and personnel skills to technology limitations. All the constraints guide ongoing conversations on the desired implementation of the technology solutions, exposing each community's hidden features and reinforcing the transparent interdependence.

As you apply ML, you'll recognize many of these boundaries as already being part of your job. The key is to put them in the right place(s) and keep them current. We can recognize boundaries by a traditional case-by-case process, too, but because we're all busy enough with the problems of applying ML to business process, we may want to rely instead on semantic technology to do some of the heavy lifting.

As in the deployment of other analytics technologies to serve the needs of an enterprise, the idea is to raise effective questions and get conclusive answers to those questions. The only difference with the addition of ML is the scope, range, and pervasiveness of those questions and answers as the organization accommodates ML.

I suggest aligned, parallel organizational processes for machine learning and for the organization, as indicated in the two pie charts of Figure 3. The center of the figure represents a complex, formalized Q&A ontology that's intermediated by a decision support system (DSS).

Bootstrapping Transparency with Stories

"Fit" and "feel" apply both to the problem and the means to solve it. Yet, in many applications things come off the rails because no one knows the right questions to ask to begin with. Ironically, with the NASA case study outlined earlier, the technology could shape the questions better than either the implementers or stakeholders because the inputs, outputs, and connecting mathematics between them fit together only in very specific ways.

Building a knowledge model of the case study's minimal model became the "transparency bootstrap." Done as an autonomously generated initial decision tree, it was a means to express the problem in the simplest form for all. Immediately, the means to address and annotate this initial decision tree with machine learning enriched the semantic model with different ML techniques. These techniques can be

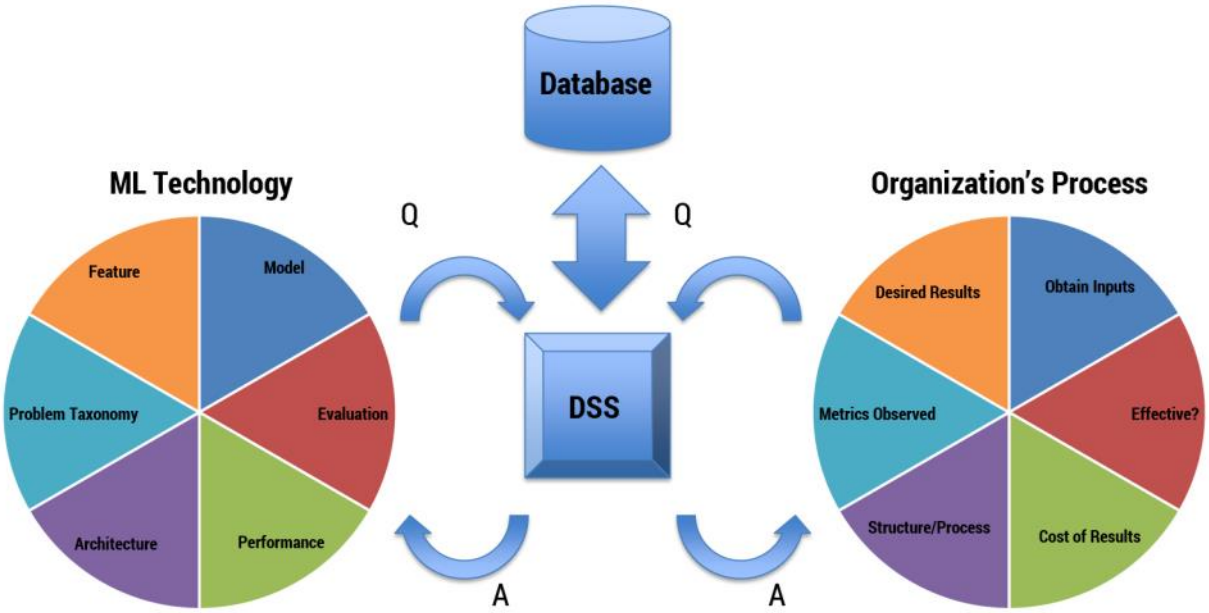

Figure 3 — Parallel ML and organizational processes with DSS/Q&A intermediation.

made "conversational" by production systems that form stories and questions about the ML techniques, again shared by all. Additions to the story either elaborate on the decision tree or constrain the bounds of the decisions.

The business community is ahead in this area because it has documented, metricized, and subjectively "knows the story" better than the technology side does. Because the business community owns the story — including contradictions and flaws that it alone can remove — it can quickly respond to adjust the narrative to fit its needs and is not bashful about doing so. It's been my experience that when a process (or definition) is explained in a story form, those telling it are the first to notice a longstanding discrepancy.

The technology side often flails to implement the changes the business side identifies as needed and may even have radically different changes to sort out the relevant factors into a more disciplined problem statement, where the resulting stories and questions are stated differently than before. The technology side is like a promising understudy, with the benefit of blissful ignorance along with the need for critical answers that decide strategies for ML applications.

The Q&A ontology is a directed graph of graphs that has at vertices either trees or another graph. It is composed of a mixture of constraints, properties, and implications. Implications are rooted by location in the graph. The path describes the context of the question where it matters, and the graph vertex (contents or descendants) describes the net meaning or implication. The point here is to ask a deep, symbolic question by placing an empty implication. By filling it, the question is answered. The answer is of the form of a union of paths in the ontology of specific other implications that are joined by connector logic (and/or/succession/exclusion). The DSS is the means to view, change, alter, notify, or trigger on a community member's domain of the ontology.

False Starts

Current practice of this Q&A ontology is erratic and, where things break down, unstructured and haphazard. The erratic nature usually arises from an opaque initial series of false starts in judging techniques, data, problem statement, efficacy, and the like. Often, such efforts are forced to conclusion without any transparency — and with a hit-or-miss result. Sometimes it is possible to "graft" a limited scope window into the Q&A as a process but to call it transparent is a bit of a stretch. Even more of a stretch are the attempts of ML vendors to "explain" such models (see sidebar, "The Desire for Technology to 'Do It All' by 'Explaining' "). Irrelevant factors dominate and distract, and only subject matter experts can remove them.

The Desire for Technology to "Do It All" by "Explaining"

One of the unsolved problems in artificial intelligence (AI) is creating a general problem solver, which inevitably fails because common terms are too subjectively multivariate. As a result, efforts at using other forms of AI to substitute for subject matter experts only multiply the effort — for no gain and additional risk. When we ask for an explanation of an ML model by an AI so we can judge impact, highly subjective terms are assigned meaning that misses subtlety — or worse, becomes self-referential and thus meaningless. Deming might have considered this a form of "A system cannot understand itself. Understanding comes from outside."[1]

Existing experts may feel they already "have this." They may view the issue as arising from "deep Q&A" between communities because they've "come from outside." Explaining adds confusion because of the need to explain nuance to a machine, while also still struggling with what the machine is attempting to communicate about the model.

Having too much AI in the picture when handling broad ML deployment could multiply the size and scope without gaining the clarity to judge transparency.

[1]Stevens, Tim. "Dr. Deming: 'Management Today Does Not Know What Its Job Is' (Part 2)." *Industry Week*, 18 January 1994.

It is presumptuous to automate insight into ML deployment as an alternative to the transparent process with Q&A that I argue for in this article. Only the implementers and stakeholders can nail down the specific meaning of the "stories" and "questions," so that the semantics in the ontology can become concrete.

Thankfully, the art of good business is all about concrete stories, questions, and the corresponding answers to those questions. The DSS Q&A ontology effectively achieves this goal in a unity of implementer and stakeholder communities. It is a realization of both communities, concrete but changeable, with full access throughout the organization. The DSS Q&A ontology is the only source of detailed joint problem definition of the ML deployment, no matter how it is described or distributed throughout an organization. Any change made across the organization impacts all affected, so the scope of ramifications matches the scope of the ML model's organizational "reach." Transparency allows stakeholders to see the "hand" of ML through this maintained (and gained) transparency, which is the benefit stakeholders seek.

Situational and Transactional "Crossing the Chasm"

Implementers and stakeholders won't be content to play in their own subject matter domains. Both sides will have a voice in any alternative choices around implementation and ownership of process. Among these choices will be alternatives that exchange accuracy, quality, and other properties for different benefits to the other side.

In addition to the need for model and feature completion in the ML application, additional constraints and behavior present in the ontology will affect model and feature completion as an interchange. This interchange has the side effect of making incrementally transparent what would otherwise be opaque constraints, behavior, and implementation details. Because they are intermediated and captured in the ontology, these details are retained intact across the entire organization.

Because all these decisions are "on the table" — visible to all and immediately changeable — anyone in the organization can see any potential impact and take responsibility for his or her domain. Even when an obscure change to an ML process or algorithm occurs, those responsible are alerted to the scope of change — to where, how, and what the change "touches" in that domain (e.g., a part's quality dynamically changes in a materials requirements planning [MRP] or enterprise resource planning [ERP] system) — so that they can supply relevant scope or constraint detail [e.g., batch or lot size or orders, specific vendors, or returns]).

In the NASA case study, this dynamic process was the means to resolving how to choose between sequential rendezvous with debris versus waiting for optimal timing (phasing) for such opportunities to occur, trading orbital maneuvers and resources for the benefit of when best to apply them. In business, many subjective and objective choices are intertwined in the problem to be solved, and some of these are "ordered" as well.

This is where both stakeholders and implementors become of one mind because these decisions can be made jointly as the transparency retained by this process sorts out the contributions of each and how they are interposed.

Normative Common Decision Making Across All — One "Thing" Only!

The final step is for joint decisions to lead to joint ownership of the ML application. All of the communities are "signed" to the same norms for common decisions, so the decision to not "compartmentalize" during broad ML adoption at the beginning pays its dividends as the unity of decision couples with ML's power to discover important details unknown to the stakeholders that were present in something previously inaccessible.

The entire point of this approach, as illustrated in the NASA case study, is to expose the total scope of a project or mission to optimization. This is made possible by carrying transparency through every step with change management. All stakeholders and implementers can accept where the ML application affects them because they were party to it from the start. It is this strength that energizes a competitive business. Everyone feels the comfort of working together while still retaining individual professionalism.

How Does This Work in Practice?

Remember I mentioned "decisions"? DSSs are the ideal means to connect ML efforts and models to the rest of the organization's existing (or being restructured) business processes, metrics, and quantitative objectives. Many already use DSSs to sift through metrics and process data to support decisions.

The twist here is that we "build" the DSS database in a special way that allows for the Q&A ontology, independent of sequence or order. At any time, the ML technology, model, metadata, or business definitions and processes can change (symbolically or numerically). When this occurs, it is as if the entire timeline has changed, and everyone's presumptions are revisited (via automated notification, alerted through email/devices). The results appear as brief contextual descriptions or "stories" in both communities' domains. One can judge the before and after as an "A/B test" to see what has occurred. No one is blindsided by a change; everyone can see how it affects them.

The surprise is that you already have hundreds of existing documents that describe all of the above. Some will be out of date, sloppy, missing a few details, and/or contain many irrelevancies. *It doesn't matter*. Since you started with the simplest "bootstrap" description, as you add, the DSS changes, propagates to both communities, and incrementally refines before you. There is no huge burden before there are results, and you can withdraw or restate when you don't like what's happening. The burden is not in the content generation, but in the choosing and inclusion.

> *As they improve with process acuity in ML adoption, human appreciation for the clarity of experience grows, as a sort of "yin and yang" to the practice of business.*

This isn't how a traditional DSS works, nor does a traditional DSS work with this kind of data and process, so the DSS database is not as "off the shelf" as it might seem. Since the entire history of operation is subject to revision, aggregate database size while tracking all of these changes has the potential to be exponential, so there can be scaling issues. But this isn't an impossible obstacle to scaling, because scaling the DSS database pervasively across an organization is tractable.

Making changes is not automated. The changes require subject matter expertise to interpret, sometimes resulting in revision of the description of the business process because a more detailed statement is necessary for a desired result. This is tedious to begin with, but one side benefit is that it makes experts even more "expert" because they get a chance to have an even more critical focus than they were allowed before.

Perhaps this foreshadows what could happen to organizations longer term. As they improve with process acuity in ML adoption, human appreciation for the clarity of experience grows, as a sort of "yin and yang" to the practice of business.

Conclusion

ML isn't new, but it is very opaque. If we want to apply it broadly to an already opaque organization to obtain the desired gains, we may need to apply it from the start in such a way as to make both the ML and business processes transparent.

This process of applying ML will not occur just once in an organization. We are already seeing technology improvements, even as pilot projects are still underway. More implementations of ML will follow, requiring continued interaction with business processes, as this nexus of specification connects even more broadly with the organization, including new pilot projects and novel ML techniques as they become available. The way in which we structure business processes alone deeply depends on considerable organizational knowledge that is rapidly changing — much more rapidly than we can keep documented.

Transparency isn't a minor virtue but a major one. In a widespread ML adoption effort, longstanding practices that weren't noticed before might suddenly matter. So, by retaining and increasing transparency, not only might the technology become more effective than it otherwise would, the organization's comfort with those now-improved, longstanding practices grows in measure. Alongside more effective technology and increased comfort, the ability to execute with scale also increases.

The value of transparency is to allow collaboration across business stakeholders so all can have a knowledgeable voice without opaque, "no go" areas. Addressing this with tools and process to have a common framework to keep the burden of terms and unanticipated impact to a minimum means the conversation doesn't grind to a halt and become the opaque bottleneck.

References

[1]The single most important skill of a data scientist has nothing to do with data science or ML or any technology. It's about the training in math to assemble definitions that can be used to express proofs of any kind: for making things tractable. Management must then back this up with policy definitions so those definitions can't change.

[2]Scudellari, Megan. "AI-Human 'Hive Mind' Diagnoses Pneumonia." *IEEE Spectrum*, 13 September 2018.

[3]Gianfrancesco, Milena A., et al. "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data." *JAMA Internal Medicine*, November 2018.

[4]Stevens, Tim. "Dr. Deming: 'Management Today Does Not Know What Its Job Is' (Part 2)." *Industry Week*, 18 January 1994.

[5]In *The Hitchhiker's Guide to the Galaxy*, the Ultimate Computer is asked to find the answer to the ambiguous question of "Life, the Universe, and all that?" to which it ponders and answers perfunctorily as "42"; see: Adams, Douglas. *The Hitchhiker's Guide to the Galaxy*. Golancz, 2012.

William Jolitz is founder of Valux and leads fundraising and portfolio management. As an investor, entrepreneur, and inventor of today's technologies, he has handled M&A for companies like Tandem Computers and due diligence for venture capital firms. Mr. Jolitz has founded, grown, and exited several Silicon Valley–based technology companies in systems, enterprise, and Internet. He is the inventor of OpenBSD (open source Berkeley Software Distribution), the progenitor of iOS, Linux, and Android. Mr. Jolitz holds a bachelor of arts degree in computer science from the University of California, Berkeley. He can be reached via LinkedIn.

NOTHING IS PERFECT

Vulnerability and Risk Mitigation in AI and Machine Learning

by David Biros, Madhav Sharma, and Jacob Biros

Artificial intelligence (AI) and machine learning (ML) offer promising technologies in areas such as healthcare, transportation, and finance. Information engineers and computer scientists, together with data analytics experts, have developed systems to monitor and control prescription medicine dosages and sequencing (in cases of multiple types of medicine);[1] have created autonomous, self-driving vehicles;[2] and have helped prevent financial crimes such as credit card fraud.[3] Indeed, AI/ML can offer a world of good. Although, as recent debacles have shown, that is not always the case. Between poor implementation, lack of understanding of the technology, and a general tendency among companies to assume AI/ML is the solution to all their woes, there are a multitude of ways that the end results can be problematic.

One such example occurred when a Washington, DC, USA, school district fired a teacher based on an algorithm's recommendations.[4] The system, designed to evaluate teachers based on their students' test scores, did not consider that the students in the district were impoverished and faced additional obstacles that prevented them from scoring favorably on the tests. The deeply committed teacher, who ranked high on other forms of evaluation such as direct observation by her superiors, is now teaching in another community.

Table 1 highlights 18 failed AI/ML projects and the reasons behind those failures. Examples like these show that AI/ML has vulnerabilities that can create serious consequences for lives and property. Our research aims to investigate the challenges associated with employing AI/ML and answer the following two questions:

1. What are the component vulnerabilities of AI/ML?
2. How can those vulnerabilities be mitigated?

We start by investigating AI's vulnerabilities, examining where problems can occur in development and application. Then we consider the impact those vulnerabilities can pose for quality decision making and offer mitigation strategies for each component. Finally, we suggest plans for mitigating risks associated with AI/ML to avoid future, similar pitfalls.

The Challenges of Working with AI

Based on the cases noted in Table 1 and using an inductive approach, we identified three major challenges that can lead to failure in AI/ML-related projects: problem fit, data, and application. In this section, we consider the risk and impact of failure for each of these three challenges.

Problem-fit issues arise from development and deployment of AI/ML when these applications are not consistent with the organization's problems, goals, or values.

Problem Fit

AI and ML are currently very popular in media and academia. Organizations are encouraged to develop AI/ML applications for nearly all new projects. Problem-fit issues arise from development and deployment of AI/ML when these applications are not consistent with the organization's problems, goals, or values. In the words of American psychologist Abraham Harold Maslow, "I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail."[5] Due to expansive commoditization in a relatively new marketplace, the stories of companies losing money with AI are not typically publicized.[6] However, the examples that are available would imply that attempts to apply AI/ML without fully understanding the technology and its impacts do exist and, in some cases, have led to potentially disastrous results.

Let's look at two examples of bad problem fit. A gender-neutral social media ad for STEM careers was seen by many more men than women.[7] Further

Table 1 — 18 failed AI/ML projects.

Company/Project	Failure and Impact	Reason
Admiral Insurance firstcarquote	Insurance company piloted a program to set rates based on algorithms' determination of consumers' character from Facebook posts and likes. Facebook blocked the insurer.	Application
Amazon AI Recruiting Tool	Recruiting tool was trained on data featuring résumés from male-dominated tech companies, making the tool show bias against women.	Data
Amazon Alexa	When a child requested that the voice assistant play a children's song, the assistant instead selected pornography.	Application
Amazon Rekognition	Amazon pitched a facial recognition tool that misclassified women and people of color 19% of the time.	Data
DC Public Schools Teacher Evaluation	School fired a teacher based on opaque algorithm that measured the teacher's performance on student test scores and did not consider humanistic factors.	Data
Faception Profiling Character Based on Facial Image	Israeli startup claims to have an accurate algorithm that uses 15 classifiers to analyze people's personality traits to determine whether a person is a pedophile, terrorist, etc. Scientists worldwide have questioned the moral and scientific merit of claims.	Application
FakeApp Deepfake	Deepfake tool uses deep learning to morph (or replace) faces or pictures, enabling production of fake videos. Tool led to tainted reputations for numerous celebrities, aided cyberbullying, and facilitated the spread of misinformation.	Application
Google Project Maven	Google partnered with the Pentagon to use AI tools to identify potential drone targets. Employee backlash led to the company's nonrenewal of contract.	Problem fit
IBM Watson for Oncology	Supercomputer provided unsafe recommendations for cancer treatment during simulations.	Problem fit
Los Angeles Police Department CalGang	LAPD's database of suspected gang members was found to contain numerous unsubstantiated and inaccurate entries.	Data
Microsoft Tay	Microsoft introduced chatbot to improve its algorithm's "conversational understanding." Chatbot was corrupted within 24 hours due to troll tweets.	Data
New Zealand Dept. of Internal Affairs Passport Picture Check	Government agency rejected a passport application from a man of Asian descent because facial recognition software incorrectly determined that his eyes were closed.	Data
Niantic Pokémon Go	Users noted that the augmented reality game had far more PokeStops and Gyms in predominantly white neighborhoods than in predominantly minority ones. The gaming company attributed the distribution to the training data.	Data
Ningbo City Police (Zhejiang, China) Facial Recognition	Facial recognition cameras incorrectly identified a famous Chinese businesswoman as a jaywalker because her picture appeared on an advertisement on the side of a bus.	Application
Police Criminal Risk Assessment Algorithm	Police use algorithm to predict likelihood to re-offend (recidivism) and present the predicted score to judge. The algorithm's data resulted in a high recidivism risk score for low-income and minority communities.	Data
Uber Self-Driving Car	A driverless car had an accident during testing in Arizona, USA, resulting in the death of a pedestrian.	Application
Waco International Beauty.AI	An AI/ML algorithm judged an international beauty contest based on supposedly objective factors such as facial symmetry and wrinkles. The algorithm's 44 winners, from the over 6,000 participants of all races, were mostly white.	Data
YouTube Recommendation Algorithm	Recommendation algorithm centered on maximizing watch time led to recommendation of objectionable and radicalizing content.	Application

analysis revealed that the gender gap was due to the algorithm's optimizing on price. Under the user-defined parameters, the algorithm was tasked with showing the ad to the maximum number of viewers possible, while minimizing cost. The demographic "Young Women" is more expensive than other demographics with respect to advertising, and the algorithm, therefore, excluded this demographic from its purchasing. Better analysis of the problem to be solved (budget constraints or wide exposure to a gender-neutral ad) potentially would have avoided the apparent gender-based discrimination.

In another example, Google's involvement in Project Maven, an AI-based solution to help the military identify potential drone targets, faced backlash from Google employees and was subsequently terminated.[8] As with many other technologies (big data, blockchain), there have been and will continue to be cases where an organization's leadership opts for AI solutions when they do not fit the problem or the goals and values of the company or its employees.

From the examples investigated in this study, the risk probability of bad problem fit appears to be low. The probability of large organizations making misinformed decisions is lower than that of new ventures due to differences in commitment of resources and dependency on raising capital. However, it must be noted that many AI failures in which the solutions do not significantly benefit the organization are never reported.

Data

One of the most common problems with AI/ML systems is the quality and representativeness of the data provided. We have identified four potential challenges related to data that can cause AI/ML systems to fail.

First, the data can suffer from *imbalance issues* (i.e., it does not represent the population it was intended to represent). Amazon developed an AI tool to assist in hiring decisions in its HR department.[9] The intent was to have the tool select the highest-quality applicants without regard to gender, age, or race. However, Amazon's algorithm was trained primarily using a data set that contained résumés of mostly male candidates. As a result of training on this imbalanced data set, the tool became biased against women and didn't recommend any female applicants for jobs. In another example, a facial recognition system of the New Zealand's Department of Internal Affairs rejected the passport application of a man of Asian descent because it determined that his eyes were closed in his submitted photo.[10] The facial recognition AI had not been sufficiently trained to evaluate the faces of people of Asian descent due to a lack of representation in the training set.

The second challenge occurs when the data used for training is *used blindly* without confirming the contents. In 2016, Microsoft introduced a chatbot named Tay to engage and entertain young adults and investigate "conversational understanding."[11] Tay was to continue to learn based on the conversations it encountered on Twitter (i.e., tweets were additional training data). The bot was corrupted within 24 hours when troll tweets coaxed it into spouting racist epithets. As is the case with most algorithms, Tay had no understanding of what is or is not appropriate and simply tried to predict correct outputs based on the data on which it had been trained. If a large portion of the training data contains inappropriate text, then most algorithms will assume that is the correct way to respond. Algorithms are only as good as the data upon which they are built.

One of the most common problems with AI/ML systems is the quality and representativeness of the data provided.

A third challenge is that data can suffer from *human biases*. For example, in the case of *Pokémon Go*, users noticed that the augmented reality (AR)-based game had far more PokeStops and Gyms in predominantly Caucasian neighborhoods than in minority-dominated neighborhoods.[12] The game company attributed the distribution to the training data, which was crowd-sourced by mostly Caucasian male, tech-savvy players. In this case, human bias from the players propagated throughout the system.

Finally, data can be *inaccurate or incomplete*, thereby skewing the outcome of the AI solution. Data collected for research purposes is often tested for validity and reliability. That may not always be the case with commercial AI tools. Furthermore, if humans manually enter data into the system without controls to ensure its accuracy and completeness, then errors are very likely to occur. In turn, the results of the AI solution will be suboptimal. For example, California's gang database has been rife with errors, with entries including babies

and minors as a result of incorrect data input.[13] If this data were to be used for analysis to predict future crimes or metrics for a legal action, the consequences could be dire.

Quality of data is often correlated with sample size. Social media companies such as Facebook and Twitter have billions of subscribers, making their data closer in representativeness to the actual population. On the other hand, law enforcement agencies building databases based on criminal indictments and convictions, or startups using a smaller sample size, may suffer from low-powered tests.[14] For organizations with limited access to data that are undertaking AI/ML projects, the probability of AI/ML projects having poor-quality data is high and thus the potential negative impact of that data on AI/ML solution results is also high.

The risk of undesirable consequences due to application issues originates from the project's conceptualization.

Application

Application issues arise when AI/ML is not ethically conceptualized, properly tested, or ready to be deployed. Barring an official audit, there is no way to know where the problem originated, but the application of AI/ML may have undesirable consequences even if the application performs as designed. For example, an Israeli startup company claims to have an AI application that can determine whether people are pedophiles or terrorists simply by examining images of a person's face and analyzing 15 classifiers.[15] Many have questioned the moral and scientific validity of the company's claims. While the study of facial characteristics or physiognomy has been around for many years, it has come under serious ethical scrutiny.[16]

Applications also suffer from inadequate testing before deployment. In Ningbo, China, facial recognition cameras were used to catch jaywalkers by identifying people who crossed the street at places other than crosswalks. Police publicly shamed a famous Chinese businesswoman for jaywalking only to learn later that the AI application had captured a picture of her that was displayed on the side of a bus and incorrectly determined she was breaking the law.[17] In a far-worse example, Uber was testing self-driving cars in Tempe, Arizona, USA, when one of the company's vehicles struck and killed a pedestrian.[18] The AI technology in these two applications clearly was deployed in public well before it was ready.

The risk of undesirable consequences due to application issues originates from the project's conceptualization. Though not all unintended consequences can be predicted in advance, proper testing can make organizations aware of potential harm an application may cause. The application issue cases we investigated in this study indicate that there is a high risk of undesirable consequences with considerable negative impact.

Mitigating Risks Associated with AI

Code of Ethics and Need for AI Ethicists

Most of the cases discussed in this article show that the use of AI can have negative social implications, such as propagation of gender and race bias, incorrect evaluation of performance, or unintended consequences due to premature deployment. One step toward mitigating risks associated with AI/ML is to develop a code of ethics for developers to follow.

The Institute of Electrical and Electronic Engineers (IEEE) and the Association for Computing Machinery (ACM) jointly published a code of ethics for software developers in 1999.[19] The code includes eight principles, with the first noting the software developers' responsibility to the public. Another section of the code states that software engineers should "approve software only if they have a well-founded belief that it is safe, meets specifications, passes appropriate tests, and does not diminish quality of life, diminish privacy or harm the environment. The ultimate effect of the work should be to the public good." AI and ML developers — a subset of software engineers — should abide by this code of ethics as well.

In addition to IEEE, major accreditation bodies such as ABET and the Association to Advance Collegiate Schools of Business (AACSB) International have highlighted the need for proper understanding of ethics as an important student outcome. ABET accreditation criteria state that students must have "an ability to recognize ethical and professional responsibilities in engineering situations and make informed judgments, which must consider the impact of engineering solutions in global, economic, environmental, and societal contexts."[20]

Many data science professionals are self-trained or trained by popular massive open online courses (MOOCs).[21] Popular MOOCs (such as the one by Andrew Ng, with over 1.7 million enrollments) also underscore the importance of ethics while developing AI/ML systems.[22] Nevertheless, the lack of formal training and the frenzy around new technology may impede developers' adherence to ethics, highlighting the need for additional oversight. In lieu of academic training, organizations working with self-trained data scientists should emphasize the significance of ethics and consider adopting a code of ethics along the lines of the one developed by IEEE and ACM.

Apart from the risk of social harm, flawed algorithms often result in public relations and legal fiascos. These factors have motivated changes among market leaders in AI. For example, both Microsoft and Salesforce have created ethics committees and a senior leadership position of "chief AI ethicist."[23] Though the position is relatively rare and lacks a proper definition, the role of an ethicist is similar to that of an auditor or quality inspector. AI/ML ethicists can vet data-evaluating AI algorithms for racial bias, gender bias, and other unintended consequences. From a data standpoint, an AI/ML ethicist could work alongside analysts to provide insight on the risks involved with working with certain types of data, evaluate data imbalances for potential ethical issues, and create hypotheses about the ways in which an AI/ML algorithm may inadvertently produce unethical results. By searching for potential ethical issues before they occur, it is possible both to improve the overall ethicality of AI/ML products as well as prevent the costly project rebuilds and brand-damaging backlashes that could foreseeably come with the discovery of an ethical issue further downstream.

Proper Training and Testing

Training and testing are important stages in the development of any AI/ML algorithm. Even an ethically conceptualized and vetted application can lead to failure if the inputs and outputs of the AI/ML system have not been properly tested. Developers should test data for its quality and representativeness to ensure validity and reliability of results. Scientific experiments have faced many challenges in addressing the validity and reliability of data, and scientists have proposed numerous techniques to assess and overcome these issues.[24] For example, a deception detection study found a 1:4 ratio split between deceptive and truthful statements.[25] To make their results more reliable and use their data efficiently, the researchers partitioned the data set into four equally matched partitions. Some other common approaches include oversampling and undersampling. In addition, using techniques such as cross-validation can help ensure that AI/ML algorithms generalize to the entire training data set rather than giving too much weight to a certain portion of it.

Testing the AI/ML system's output is equally important for ensuring its success. After investing US $62 million in "Watson for Oncology," IBM had to shelve the project when, in hypothetical cases, the system repeatedly gave unsafe recommendations for cancer treatment.[26] Even though IBM incurred high financial loss, advanced stages of testing ensured no harm was caused to actual patients. Similar testing should be conducted for algorithms used in contexts with comparatively lower stakes to mitigate unintended consequences.

AI built using artificial neural networks diffuses information in a way that is difficult to decipher, and the logic behind the decisions taken by the algorithm is often based on rules that are not easily interpretable.

Vigilance

In the 1980s and 1990s, as automation became increasingly common, considerable research was conducted on trust in computers[27] and the need for human vigilance.[28] At the time, the concern was that humans were placing too much trust in computers and automation, which could lead to disastrous results. Now, while AI/ML technology is in its infancy but gaining momentum at a very rapid pace, it is prudent to issue a similar warning about over-trust and lack of sufficient vigilance.

Many AI/ML systems are continually being retrained on new data to improve the systems' overall accuracy as well as ensure that the model results continue to be relevant. It is important not only to confirm the quality of the data when an AI/ML system first begins development, but also as the system is subsequently retrained. In the case of Microsoft's Tay, the system was continuously being retrained on the data provided by people that conversed with it. Unfortunately, no measures were in place to confirm that the data it was consuming as it conversed with the public was appropriate,

and unscrupulous users were introducing it to new, undesirable data that resulted in the system behaving in an inappropriate manner. The developers mistakenly assumed users would communicate with the system in the same way the developers did and did not monitor the outcome. Their trust and lack of vigilance contributed to Tay's failure.

A major obstacle to vigilance in AI is the "black box" issue.[29] AI built using artificial neural networks diffuses information in a way that is difficult to decipher, and the logic behind the decisions taken by the algorithm is often based on rules that are not easily interpretable. The answer to this problem may be to not accept a black box approach and move away from it. Research and industry developers are starting to recognize and address this issue by increasing their understanding about the proper use of opaque programs.

As with other systems development initiatives, AI/ML projects should be rigorously tested to increase the transparency and help developers better understand the inner workings. Testing techniques such as code inspection, walkthroughs, desk checking, unit testing, integration testing, and system testing must be employed to ensure more clarity with respect to AI/ML operations.[30] In addition, there are techniques for understanding how an algorithm is making decisions. For instance, decision trees can be printed out to see how decisions are being made at every level.

We implore software engineers, despite their rush to get AI/ML applications out to market, to follow sound systems analysis and design principles, especially in employing due diligence in systems and algorithm testing. Doing so will allow them to do away, in part, with the black box mentality and help avoid potentially disastrous outcomes.

By their very nature AI/ML solutions will not give perfect answers 100% of the time, and they therefore require oversight. They must be strictly monitored until developers have a general understanding of how the solution behaves and continuously tracked to assure continued performance and to scrutinize potential outlier cases. Developers and evaluators must be ready to intervene as needed. In addition, developers should carefully evaluate and review feedback from parties involved with the system to ensure the system is performing as desired.

Conclusion

AI and ML are indeed technologies that can solve problems and do a world of good. But we must not forget that they are fallible and require human oversight. There are now dozens of documented cases where AI/ML applications have resulted in suboptimal and sometimes disastrous results. AI/ML developers must adhere to a strong code of ethics and employ ethicists to audit new applications. They must also ensure valid and reliable data testing and remain vigilant to the possibility that the technology may not work as intended. Organizations take risks when they develop and deploy AI/ML, but if these warnings are heeded, the technology can be developed safely and the risks can be mitigated to an acceptable level.

References

[1]Davazdahemami, Behrooz, and Dursun Delen. "Examining the Effect of Prescription Sequence on Developing Adverse Drug Reactions; The Case of Renal Failure in Diabetic Patients." *International Journal of Medical Informatics*, Vol. 125, May 2019.

[2]Ozbay, Kaan, Xuegang (Jeff) Ban, and C.Y. David Yang. "Developments in Connected and Automated Vehicles." *Journal of Intelligent Transportation Systems*, Vol. 22, No. 3, June 2018.

[3]West, Jarrod, and Maumita Bhattacharya. "Intelligent Financial Fraud Detection: A Comprehensive Review." *Computers & Security*, Vol. 57, March 2016.

[4]McGuire, Pat. "Death by Algorithm." *Huffington Post*, 8 May 2012.

[5]Maslow, Abraham Harold. *The Psychology of Science: A Reconnaissance*. Harper & Row, 1966.

[6]Davenport, Thomas H. "Will AI Companies Make Any Money." *Harvard Business Review*, 12 July 2016.

[7]Lambrecht, Anja, and Catherine E. Tucker. "Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads." *Management Science*, Vol. 65, No. 7, July 2019.

[8]Griffith, Erin. "Google Won't Renew Controversial Pentagon AI Project." *Wired*, 1 June 2018.

[9]Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." Reuters, 9 October 2018.

[10]"New Zealand Passport Robot Tells Applicant of Asian Descent to Open Eyes." Reuters, 7 December 2016.

[11]Kastrenakes, Jacob. "Microsoft Made a Chatbot That Tweets Like a Teen." The Verge, 23 March 2016.

[12]Akhtar, Allana. "Is Pokémon Go Racist? How the App May Be Redlining Communities of Color." *USA Today*, 9 August 2016.

[13]Winton, Richard. "California Gang Database Plagued with Errors, Unsubstantiated Entries, State Auditor Finds." *Los Angeles Times*, 11 August 2016.

[14]Shrout, Patrick E., and Joseph L. Rodgers. "Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis." *Annual Review of Psychology*, Vol. 69, January 2018.

[15]McFarland, Matt. "Terrorist or Pedophile? This Start-Up Says It Can Out Secrets by Analyzing Faces." *The Washington Post*, 24 May 2016.

[16]Agüera y Arcas, Blaise, Alexander Todorov, and Margaret Mitchell. "Do Algorithms Reveal Sexual Orientation or Just Expose Our Stereotypes?" Medium, 11 January 2018.

[17]Dodds, Laurence. "Chinese Businesswoman Accused of Jaywalking After AI Camera Spots Her Face on an Advert." *The Telegraph*, 25 November 2018.

[18]Wakabayashi, Daisuke. "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam." *The New York Times*, 19 March 2018.

[19]"Code of Ethics." The Institute of Electrical and Electronic Engineers (IEEE) – Computer Society (CS)/Association for Computing Machinery (ACM) Joint Task Force on Software Engineering Ethics and Professional Practices, 1999.

[20]"Criteria for Accrediting Engineering Programs, 2018–2019." ABET, 2017.

[21]Rayome, Alison DeNisco. "Report: 59% of Employed Data Scientists Learned Skills on Their Own or via a MOOC." TechRepublic, 30 October 2017.

[22]Young, Jeffrey R. "Andrew Ng Is Probably Teaching More Students Than Anyone Else on the Planet. (Without a University Involved." EdSurge, 7 June 2018.

[23]Murawski, John. "Need for AI Ethicists Becomes Clearer as Companies Admit Tech's Flaws." *The Wall Street Journal*, 1 March 2019.

[24]Carmines, Edward G., and Richard A. Zeller. *Reliability and Validity Assessment*. Sage Publications, 1979.

[25]Fuller, Christie M., David P. Biros, and Rick L. Wilson. "Decision Support for Determining Veracity via Linguistic-Based Cues." *Decision Support Systems*, Vol. 46, No. 3, 2009.

[26]Chen, Angela. "IBM's Watson Gave Unsafe Recommendations for Treating Cancer." The Verge, 26 July 2018.

[27]Muir, Bonnie M. "Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems." *Ergonomics*, Vol. 37, No. 11, 1994.

[28]Parasuraman, Raja. "Human-Computer Monitoring." *Human Factors*, Vol. 29, No. 6, December 1987.

[29]Castelvecchi, Davide. "Can We Open the Black Box of AI?" *Nature*, Vol. 538, 5 October 2016.

[30]George, Joey F., and Joseph S. Valacich. *Modern System Analysis and Design*. 8th edition. Pearson, 2014.

David Biros is Associate Professor of Management Science and Information Systems and Fleming Chair of Information Technology Management at Oklahoma State University. A retired Lieutenant Colonel of the US Air Force, Dr. Biros's last assignment was as Chief Information Assurance Officer for the AF-CIO. His research interests include deception detection, insider threat, information system trust, and ethics in information technology. Dr. Biros has been published in MIS Quarterly, Journal of Management Information Systems, Decision Support Systems, Group Decision and Negotiation, MISQ Executive, Journal of Digital Forensics Security and Law, *and other journals and conference proceedings. He earned a master's degree in public administration from Troy State University, a master's degree in information resource management from the Air Force Institute of Technology, and a PhD in information and management sciences from Florida State University. He can be reached at david.biros@okstate.edu.*

Madhav Sharma is a PhD student, studying management science and information systems at Oklahoma State University. His research interests include diffusion of innovation, use and implication of artificial intelligence, machine learning, and the Internet of Things. Mr. Sharma earned a master's degree in telecom management and an MBA, both from Oklahoma State University. He can be reached at madhav.sharma@okstate.edu.

Jacob Biros is a mechanical engineer and data analyst at Chura Data in Okinawa, Japan. He is an experienced system developer who has worked on a wide variety of artificial intelligence/machine learning–related projects ranging from natural language processing to dynamic pricing. Mr. Biros earned a bachelor of engineering degree from Oklahoma State University. He can be reached at jakebiros@gmail.com.

PSST … HEY YOU!

When AI Nudging Goes Wrong

by Richard Veryard

Nudge theory suggests that people can be persuaded to take certain actions, or avoid other actions, via a well-chosen nudge. In many contexts, people consider persuasion better than outright coercion, since it may achieve policy objectives without limiting free choice. However, there are various ways in which nudges do not produce the expected effect and may even be counterproductive.

While there may be issues associated with any form of nudging, there are some additional concerns arising from the recent explosion in technologically mediated nudging, which may use artificial intelligence (AI) to *detect* situations where a nudge may be appropriate, to *deliver* an appropriately customized nudge to the receiving party, and, if possible, to *discover* the effect on the nudgee.

Influencing Choices

People are becoming aware of the ways in which AI and big data can be used to influence people, in accordance with nudge theory. Not only can individuals be nudged to behave in particular ways, even large-scale social systems (including elections and markets) can apparently be influenced.

While early forms of nudge theory can be found in cybernetics, it was further developed and popularized by behavioral economist Richard Thaler and legal scholar Cass Sunstein.[1] Nudge theory has been widely adopted by business organizations, governments, and other agencies, and underpins a variety of methods aiming to influence the choices of consumers and citizens. Nudges typically work by framing a person's choices in particular ways; for example, by drawing attention to options that might otherwise have been overlooked.

Although some forms of nudge may be regarded as ethically dubious, many nudges can be justified if they are well conceived, if it can be shown that they benefit (or at least do not harm) the people who are nudged, and if they are reasonably open and transparent.

Types of Nudges

In our daily lives, we may be subject to a wide range of nudges of different types and from different sources. While it is not the intention of this article to provide a full taxonomy of nudges, we can identify two significant dimensions, as shown in Figure 1 and explained further in Example 1.

First, let's consider the *mediation* dimension. This implies a separation between the *design* of the nudge and its *delivery*, regardless of whether the nudge is actually delivered by a human being or a machine. Such nudges are designed for large-scale deployment, and their effects may therefore be much broader than an individual unmediated nudge. There will typically be some form of instruction or message that triggers the delivery of the nudge, and this is, at least in principle, open to monitoring and audit.

Second, consider the *technology* dimension. While technology may not fundamentally change the nature of the nudging, it typically amplifies the reach, richness, agility, and reliability of the nudging.[2] Not only can nudges be broadcast to large numbers of people, but each nudgee can receive a nudge that is personalized for greater effect, using big data and feedback to adjust the wording, timing, and force[3] of the nudge for each recipient. There are many contexts, such as online shopping, in which a person is presented with personalized recommendations, based on his or her previous actions or other available data, which may nudge that shopper to consume the recommended items.

Advanced forms of technology would be able to design and deliver nudges autonomously, using black box algorithms and machine learning (ML). This might be difficult if not impossible to inspect and understand. These autonomous nudges bring us back to the unmediated type (refer back to Figure 1).

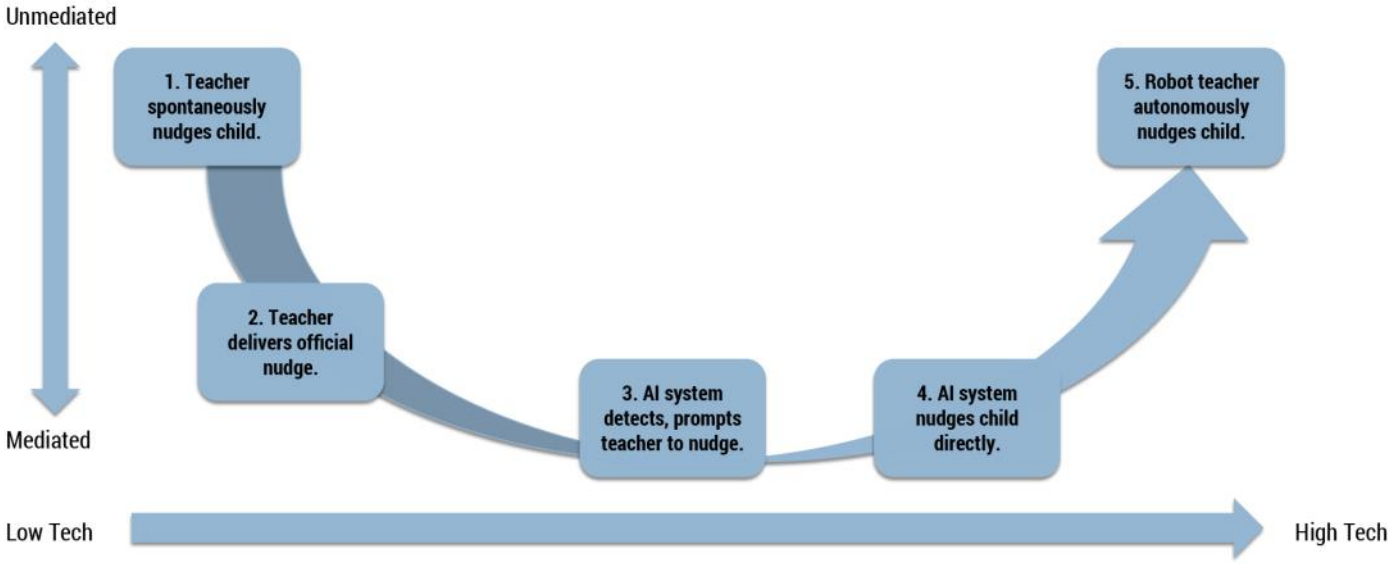

Figure 1 — Range of nudging styles: a teacher-child example.

Example 1 — Teacher-Child Nudging

1. A teacher autonomously nudges a child with no technological help. This is the simplest unmediated case.
2. A teacher nudges a child according to an official policy, procedure, or script. Now the teacher is no longer an autonomous agent delivering an unmediated nudge but is acting as the delivery channel for a nudge designed elsewhere — therefore mediated.
3. An AI system detects that the child is not paying attention and notifies the teacher, who nudges the child.
4. The AI system nudges the child directly without involving the teacher, following some centrally designed program.
5. A superintelligent and autonomous robot teacher conceives and delivers a nudge. To the extent that the robot has created the nudge itself, the robot is no longer merely the delivery channel for a nudge designed elsewhere but is equivalent to the teacher in the simple unmediated case.

The Architecture of the Nudging System

Technologically mediated nudges are delivered by a sociotechnical system we could call a *nudge system*. This system might contain several different algorithms and other components and may even have a human in the loop. Our primary concern here is about the system as a whole.

Typically, a nudge system would perform several related activities, such as the following four:

1. There would be some mechanism for "reading" the situation. For example, the system would be able to detect the events that might trigger a nudge, as well as determine the context. This might be a simple sense-and-respond mechanism, or it might include some more sophisticated analysis, using some kind of model. There is typically an element of surveillance here.
2. Assuming that there was some variety in the nudges the system produced, there would be a mechanism for selecting or constructing a specific nudge, using a set of predefined nudges or nudge templates.
3. There would then be a mechanism for delivering or otherwise executing the nudge. We might call this the *nudge agent*. In some cases, the nudge may be

delivered by a human but prompted by an intelligent system. If the nudge is publicly visible, this could allow other people to infer the circumstances leading to the nudge — therefore introducing a potential breach of privacy.

4. In some cases, there might be a direct feedback loop, giving the system immediate data about the human response to the nudge. Obviously, this will not always be possible. Nevertheless, we would expect the system to retain a record of the delivered nudges for future analysis. To support multiple feedback tempos,[4] there could be multiple components performing the feedback and learning function. Typically, the faster loops would be completely automated (autonomic), while the slower loops would have some human interaction.

There would typically be algorithms to support each of these activities, possibly based on some form of ML, and there is the potential for algorithmic bias at several points in the design of the system, as well as various forms of inaccuracy. (See Example 2, which describes a public anti-smoking nudging system.)

In many cases, there will be a separation between the technology engineers who build systems and components and the social engineers who use these systems and components to produce some commercial or social effect.

Unintended Consequences

There are several ways that a nudging system may fail to produce the intended outcomes, such as:

1. **Wrong audience.** Nudges affect people other than those targeted.

2. **Bias.** A nudging system is composed of multiple intelligent components. Each component may contain various forms of bias, and the components may interact in unpredictable ways.

3. **Resistance.** If the nudgee consciously or unconsciously resists a nudge, the nudge may trigger an action in the direction opposite to the apparent direction of the nudge.

4. **Interference.** Multiple nudges (either repeated nudges from the same system or different nudges from different systems) may have an unpredictable cumulative effect.

5. **Brutalization.** If people become accustomed to treating anthropomorphic devices such as robots in certain ways, they may transfer these behaviors to humans or animals.

6. **Manipulation.** A nudging system can be manipulated in several ways by external agents, either by misdirection of the system itself or by the creation of interfering nudges.

In some cases, there may be a suspicion that these consequences are not truly unintended and that the designers and users of the nudging system have hidden intentions that are distinct from the espoused intentions. However, for many purposes, we may be able to take their espoused intentions at face value.

Example 2 — Public Nudging System: Anti-Smoking Messages

Consider a digital advertisement in a public place that shows anti-smoking messages whenever it detects tobacco smoke. The system can distinguish different brands of cigarette and can estimate how many people are smoking in its vicinity. The system generates different anti-smoking messages for the expensive brands versus the cheap brands. The system interconnects with other systems containing personal data to identify smokers. This allows it to send messages to smokers' phones, as well as name and shame smokers on the public display board. Or, it might tell your friends and family that you were having a sneaky cigarette when you had told them that you had given up smoking.

Note that there are various opportunities for error — including false positives when the system incorrectly detects tobacco smoke or mistakes a seated adult for a child. Note also that more information doesn't always mean better information. If the system included a sensor that would estimate the height of a smoker in order to detect underage smokers, for example, this would introduce new possibilities of error.

1. Wrong Audience

Nudges may be designed to have a desired effect on a specific target group, but the effect on people outside the target group may be unwanted or unknown, as shown in Example 3. For a technological example, consider the potential errors outlined in Example 2.

2. Bias

Although the question of algorithmic bias is often raised as a concern, this is not the whole story. As indicated earlier, a nudging system is composed of multiple intelligent (algorithmic) components. Each component may contain various forms of bias, and the components may interact in unpredictable ways. There may be bias in the way the nudges are worded, and such bias may be located not in the algorithms themselves but in the templates. And if there is a human in the loop, this person's bias may also affect the process.

We may also note that if there is any bias, it may either be inherent in the design of the nudge technology itself or it may be introduced by the users of the nudge technology when customizing it for a specific purpose. For example, nudges may be deliberately labeled as "dog whistles" — designed to have a strong effect on some subjects while being ignored by others — and this can produce significant and possibly unethical bias in the working of the system. The most important question is whether the nudging system as a whole is biased — either in the way that it selects people to be nudged, in the way that specific nudges are triggered, or in the way that the nudge is constructed and delivered.

Example 3 — Persuading Smokers to Use E-Cigarettes

A nudging campaign is designed to persuade adult smokers to switch to e-cigarettes. This campaign might be justified by the argument that it is not harmful for this target group and could be beneficial, based on two assumptions: (1) e-cigarettes are safer than traditional cigarettes, and (2) many smokers will find it easier to switch to e-cigarettes than to give up smoking altogether. If this campaign reaches non-smokers, ex-smokers, or children, however, it may persuade some of them to take up e-cigarettes. This cannot be justified using the same argument.

Although various forms of imperfection in a socio-technical system might be attributed to bias, the most troubling from an ethical point of view is when this bias produces some form of injustice (see Example 4). In other words, if the nudge is regarded as beneficial to the nudgee, then it may be unfair to provide this benefit to some people and to withhold it from others. Therefore, we need to be particularly alert to any bias in the algorithm selecting which people are to be nudged, or any bias in the wording of the nudge that makes it more or less effective for different groups.

3. Resistance

If the nudgee consciously or unconsciously resists a nudge, the nudge may trigger an action in the opposite

Example 4 — Systemic Bias Producing Unfair Outcomes

1. Some people receive nudges to warn them before they receive some disciplinary or financial penalty, while others are left to incur the penalty without a warning.
2. A teaching package that nudges boys to do better at math but does not provide the same encouragement to girls creates an unbalanced offering.
3. Fitting expensive cars with systems that warn drivers of speed traps disadvantages drivers of less expensive cars who do not receive such warnings.
4. A seaside town in Southern Spain had parking meters with instructions in Spanish and English. The intent was to make parking for nonresidents more expensive than for residents. An additional instruction for registered residents, only in Spanish, reminded residents to press a button before inserting the coins in order to get a cheaper rate. This disadvantaged those long-term residents who didn't speak Spanish. (My father, who told me this story, thought it served them right for failing to learn the local language.)

direction to the apparent direction of the nudge, as shown in Example 5. In some cases, nudges may be designed with this intention, but the phenomenon of reverse psychology may also explain some unintended consequences.

Paradoxical injunctions[5] make perfect sense in terms of systems theory, which teaches us that the links from cause to effect are often complex and nonlinear. Sometimes an accumulation of positive nudges can tip a system into chaos or catastrophe.[6]

4. Interference

When a person is nudged multiple times — whether by repeated nudges from the same system over an extended period or by different nudges from various systems — the cumulative effect may be unpredictable and even counterproductive. (This is a similar problem to understanding and managing the potential interactions/interference among multiple prescription drugs, given that it is impossible to put every conceivable combination through clinical trials).

Example 5 — Propaganda

Shortly before the 2016 Brexit Referendum, an English journalist noted a proliferation of nudges trying to persuade people to vote for the UK to remain in the European Union (EU), which he labeled as "propaganda."[1] While the result of the referendum was undoubtedly affected by covert nudges in all directions, it is also easy to believe that the pro-establishment style of the "remain" nudges could have been counterproductive.

[1]Gerrans, Sam. "Propaganda Techniques Nudging UK to Remain in Europe." RT, 22 May 2016.

Example 6 — Chatbot Encourages Gross Behavior

A chatbot with a female persona responds in a flirty and submissive manner to gross questions from male users, thus encouraging and reinforcing this behavior — not only toward the chatbot but also toward humans.

Moreover, the nudgee may become overdependent on being nudged, thereby losing some element of self-control or delayed gratification. A succession of nudges that alter people's preferences, goals, or political opinions can have a significant effect, not only on the individual but also on our democratic institutions.[7]

5. Brutalization

If users get in the habit of treating anthropomorphic devices such as robots in a casual or even aggressive manner, then this may nudge them into abusive behavior to humans (see Example 6). (This relates to Immanuel Kant's notion of brutalization.) Furthermore, brutalization at the social level may damage democracy.[8]

6. Manipulation

A nudging system can be manipulated in several ways by external agents, either by misdirection of the system itself or by the creation of interfering nudges. A third party can manipulate an AI system in a variety of ways — from adversarial examples to poisoning the model.

Certainly, nudge technologies could be exploited by third parties with a commercial or political intent. For example, there are constant attempts to trick or subvert the search and recommendation algorithms used by the large platforms, and Google appears to have an ongoing battle to combat misinformation and promotion of extreme content.[9]

Toward Robust and Responsible Nudging

To avoid these consequences, creators of nudging systems should consider the planning, design and testing, and operation of the system.

First, the planning. If a nudging system is intended to produce a specific set of behaviors, then these behaviors must be clearly defined so we know what success looks like. Furthermore, the scope must be clearly defined, including target audience and context.

Second, the selection of the nudging technology and the design of the system must be transparent to the appropriate stakeholders. At a very minimum, those responsible for the nudges must understand and control

how the system works. And for meaningful consent and free choice, the recipients of the nudges should be conscious of being nudged, and the implications of the nudge.

Within technology ethics, *transparency* is a major topic. If a robot is programmed to include a predictive model of human psychology that enables it to anticipate the human response in certain situations, this model should be open to scrutiny. Although such models can easily be wrong or misguided, especially if the training data set reflects an existing bias, with reasonable levels of transparency (at least for the appropriate stakeholders), it may be easier to detect and correct these errors than to fix human misconceptions and prejudices.

Reviewing a design before it goes live may help eliminate some kinds of error or bias. For example, a more diverse review board might have picked up the potential problem with the chatbot identified in Example 6.

Third, the system should be tested on a sample of the target population to verify the planning assumptions and to determine the appropriate strength and frequency of nudges.

Finally, there needs to be some mechanism for monitoring the effects on the population at large, so that any adverse or surprising effects can be reported and investigated, leading to any necessary remedial action or learning. Among other things, this may help capture any residual brutalization or other longer-term consequences.

Conclusion

This article examined some of the unintended effects of a nudging campaign. Such effects can be caused by poor design of the campaign itself or by unexpected interaction with other events. Unintended effects can also result from flaws in the nudging tools. Therefore, it is important to test large-scale nudging campaigns properly and monitor them carefully.

References

[1]Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, 2008.

[2]Karen Yeung has introduced the concept of the hypernudge, which combines three qualities: nimble, unobtrusive, and highly potent; see: "'Hypernudge': Big Data as a Mode of Regulation by Design." *Information, Communication & Society*, Vol. 20, No. 1, 2016.

[3]If we regard the nudge as a speech act, then the strength of the nudge is related to what philosophers call *illocutionary force*.

[4]Veryard, Richard. "The Emergence of Organizational Intelligence." Cutter Consortium Data Analytics & Digital Technologies *Executive Report*, Vol. 10, No. 7, 2010.

[5]Bordenave, Richard. "When Paradoxes Inspire Nudges." *Richard B-Blog*, 6 April 2019.

[6]Donella Meadows makes this point in a classic article; see: "Leverage Points: Places to Intervene in a System." The Sustainability Institute, 1999.

[7]Zeynep Tufecki has argued this in relation to YouTube: "YouTube, the Great Radicalizer." *The New York Times*, 10 March 2018.

[8]Helbing, Dirk, et al. "Will Democracy Survive Big Data and Artificial Intelligence?" *Scientific American*, 25 February 2017.

[9]Hern, Alex. "Google Tweaked Algorithm After Rise in US Shootings." *The Guardian*, 2 July 2019.

Richard Veryard is a technology consultant specializing in the data-driven business, with a keen interest in technology ethics. His books include Component-Based Business *and* Building Organizational Intelligence. *Mr. Veryard is currently writing a book on responsibility by design. He can be reached at richard@veryard.com.*

FEAR OF MISSING OUT

Strategic Perspectives on AI Product Development

by Pavankumar Mulgund and Sam Marrazzo

It has been widely reported that the lack of an artificial intelligence (AI) strategy presents the most significant hurdle to the implementation of AI initiatives in any organization. While most companies appreciate the need for a company-wide AI strategy, they often struggle to articulate a well-thought-out future course of action that takes various strategic considerations into account. Development of such an actionable plan is difficult as it depends on a variety of contextual factors that are unique to every organization, such as the organization's digital maturity, executive buy-in, the skill and experience of team members, and resistance to change.

While AI technologies have captured the attention of computer scientists and academics for over half a century, there is now an unprecedented commercial interest in AI technologies.

Yet many companies, digital and traditional, are adopting AI technologies, driven mostly by the fear of missing out. Consequently, many investments in AI have been made as one-off implementations led by a visionary team leader, as opposed to a systematic and organization-wide strategy, although this trend seems to be changing.[1] Another challenge to the articulation of AI strategy is the emergence of a plethora of new terminology and buzzwords with no consensus as to their meanings.[2] Moreover, there is a considerable dearth of talent in the AI space. Plus, there is a minimal supply of leaders who can translate organizational business priorities into a roadmap for AI initiatives and facilitate a conducive environment suitable for the successful execution of those initiatives.

In this article, we address some of these salient issues by providing a framework for developing an AI strategy. At the outset, we define a handful of the essential elements of AI strategy. Next, we explain our framework, which not only identifies and discusses critical success factors of an AI strategy but also presents a traceability map between business objectives and corresponding technology-related decisions. Following that, we present the gradual process of developing AI capabilities within an organization and highlight some of the inevitable tradeoffs in the real-world context. Finally, we draw heavily from the cutting-edge ideas and perspectives of thought leaders in the field to provide readers with a shortlist of best practices.[3]

AI: What Is It? Why Now? How Big Is the Opportunity?

AI has been around for more than half a century. A team of illustrious scientists and engineers first coined the term in 1956, and their intended meaning was "to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."[4] Contemporary AI has undergone a considerable transformation and includes machine learning (statistical machine learning and deep learning), symbolic learning (computer speech recognition and vision), and robotics. These underlying methods, along with the emergence of sophisticated hardware components, have not only helped achieve human-like decision-making ability but also the capability to carry out a variety of complex operations. Naveen G. Rao, director of Intel's AI Products Group, notes that "we are now entering the age of AI, a time when machines are using algorithms that give them superhuman abilities."[5]

While AI technologies have captured the attention of computer scientists and academics for over half a century, there is now an unprecedented commercial interest in AI technologies. Several factors have influenced this exceptional interest. First, technological advancements in data storage and cloud computing, along with technologies like Hadoop for data mining, have increased the availability of data. In addition, with the emergence of graphics processing units (GPUs), the hardware is now available to process large volumes of

data on a real-time basis from sources such as streaming videos, machine learning (ML) data sets, and various Internet-based wireless sensors. It has therefore become viable for organizations to pursue commercial products leveraging AI technologies. Market researchers are estimating today's AI opportunity to reach nearly $78 billion by 2022.[6] With such immense market potential, many industry leaders tout AI as enabling the next wave of innovation, similar to the role of the Internet in the early 2000s.

Many companies are keen on being at the forefront and capitalizing on the opportunity. Therefore, organizations, small and large, are making massive investments in the AI space. However, as with all prior technology bubbles, only a small subset of companies will make a substantial profit from these investments. For the rest, it will be an expensive strategic error at best or an existential threat in the worst case. In what follows, we discuss the first steps in terms of strategy and implementation to make the most of the AI opportunity.

What Is AI Strategy? What Are Its Fundamental Components?

The term "strategy" is defined as a high-level plan that an organization follows to achieve sustained competitive advantage. Therefore, AI strategy, at its core, must address vital questions, such as the following: How can AI deliver better value to customers? How can it help companies increase revenues, enhance efficiency, and reduce human errors? How can AI capabilities be integrated into the existing organizational processes to develop a distinct competitive advantage?

To address these questions, AI strategy must closely align with a company's business objectives, ensuring synergy between the corporate strategy and the AI strategy. As illustrated in Figure 1, the strategy development framework begins with executive leadership identifying various strategic business outcomes. Increased revenue, reduced costs, and improved product performance are some examples of business value measures. Next, both technical experts and business leadership collectively prioritize digital and traditional process flows that have the highest potential to positively impact business outcomes. Thus, business leaders should identify workflows and use cases that bring the highest business value to the company with the least investment. The strategy team should make a quick evaluation of these use cases in terms of ROI to shortlist early candidates for implementation. Subsequently, both technical and business leaders should explore how they can embed AI within current digital and traditional workflows. While AI can augment some processes easily, others might require more disruptive transformation.

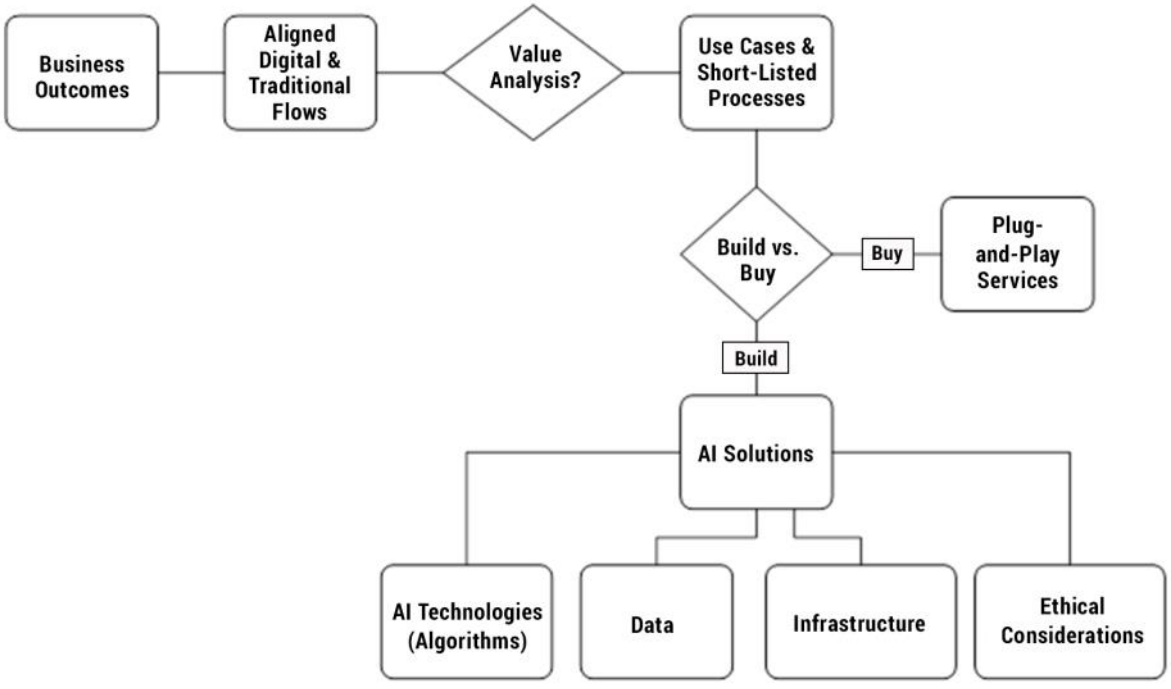

Figure 1 — Framework for AI strategy development.

Another vital consideration in AI strategy is the issue of build versus buy. The leadership and technology teams should carefully review whether they should build a comprehensive on-premise AI platform or buy plug-and-play AI services from other platforms. It is prudent to leverage the AI platforms of other companies, especially in the early phase of AI adoption, as it reduces the need for enormous up-front investment and generates quick wins that can be instrumental in acquiring executive support. Typically, companies develop a minimum viable architecture using third-party vendors such as Microsoft, Google, and IBM. As the organization's digital maturity increases, however, it should revisit the need for an AI platform. Moreover, while the organization should develop AI capabilities incrementally, it should produce an enterprise-wide vision for AI as early as possible.

In addition to alignment with business objectives, an AI strategy must consider four critical components: data, infrastructure, algorithms, and ethical considerations.

The leadership and technology teams should carefully review whether they should build a comprehensive on-premise AI platform or buy plug-and-play AI services from other platforms.

Data

Data is essential for the success of AI initiatives and is more powerful than the underlying algorithms that use the data. Companies typically perform data collection, cleansing, and preprocessing activities to ensure high data quality. However, these activities are challenging, as the data is usually fragmented across different silos, is inconsistent in format, may contain noise, and may be missing values. The digital industry has witnessed several examples where the poor quality of data has adversely affected some very sophisticated applications. One such example is that of Microsoft's Tay,[7] a conversational bot whose training continued with Twitter data, leading it to post offensive and racial tweets. It had to be shut down after only 16 hours of service. Clearly, AI strategy must carefully articulate how companies will address these data management and design challenges. Note that some data-related problems manifest from more systemic organizational issues, such as lack of cross-functional communication or the presence of an hierarchical structure. In such cases, AI adoption may require considerable organizational change.

Infrastructure

AI strategy should also consider infrastructure-related issues. Specialized hardware and software may be required to run AI and ML models effectively. Depending on the level of sensitivity of the data, organizations may decide to have an on-premise data infrastructure. For instance, some hospitals and hedge fund management companies are still averse to on-cloud infrastructure. While this option may provide the most control of the data and computing assets, it places the burden of architecting, developing, protecting, and managing infrastructure assets on the organization. Therefore, unless the nature of the data is very sensitive or the company intends to build new businesses that are entirely driven by AI capabilities, it may not be necessary to make such massive investments in infrastructure.

Algorithms

The most technical part of an AI strategy is the selection and design of the AI algorithm. However, the choice of AI algorithms depends significantly on the given use case. Most AI product development organizations build multiple models leveraging different algorithms before choosing the optimal solution.

From a theoretical standpoint, AI algorithms can either be classical or data-driven. Traditional algorithms follow a rule-based approach with deterministic outcomes, while data-driven algorithms have a range of possibilities with probabilistic outcomes. In other words, while classical algorithms produce the same result every time, the ML algorithms could produce different results for the same input depending on the data used for training. Such algorithms learn from the data and typically acquire all the biases present in the data. Over time, as AI algorithms evolve to become more sophisticated, they may also generate bias due to issues of overfitting. Addressing such concerns is among the most significant challenges in AI product development.

Ethical Considerations

Several ethical issues arise during the development of AI products. Customer privacy–related issues, job

displacement, inequitable distribution of wealth, and unfair decisions due to bias in algorithms are some notable examples. While new regulations are being created to deal with these emerging issues, organizations should place customer interests and well-being ahead of the company's interest.

How Should We Pursue the Implementation of AI Initiatives?

Most experts advise adopting a "minimum viable model" approach to the implementation of AI initiatives. Such a paradigm can be applied to both the development of the underlying AI/ML models and the platform on which these models run. For instance, if an organization is in the business of forecasting home prices, it could develop a baseline model using multiple regression. This model may have slightly higher error rates than is ideal, but in-the-ballpark forecasts are available. In the next iteration, the model can be improved using nonlinear regression. Subsequently, time series analysis can be performed to enhance the model even more. This approach ensures that the baseline model begins to deliver business value as soon as possible. On successful validation of the use case from a business perspective, the model is continuously refined to produce the best results.

Similarly, organizations should shy away from building massive infrastructure, comprising sophisticated hardware (e.g., GPUs) and platforms such as data lakes, before establishing the use case. Instead, they should follow a minimum viable model regarding platform. In this approach, they build their product on a plug-and-play AI platform. They can also leverage cloud environments such as Amazon AWS, Microsoft Azure, or Google Cloud to create AI services. After establishing the use case and validating the market need for their AI initiative, they can make more investments in infrastructure.

The gradual approach to the adoption and implementation of AI initiatives makes the most practical sense. As Figure 2 shows, some AI initiatives can be implemented quickly by integrating with AI services of a third-party platform, while others require long cycles of development, training with data, and comprehensive testing. Chatbots, text-to-speech, and automation of rudimentary tasks are some instances of low-hanging fruit that can generate some quick wins. Organizations can then

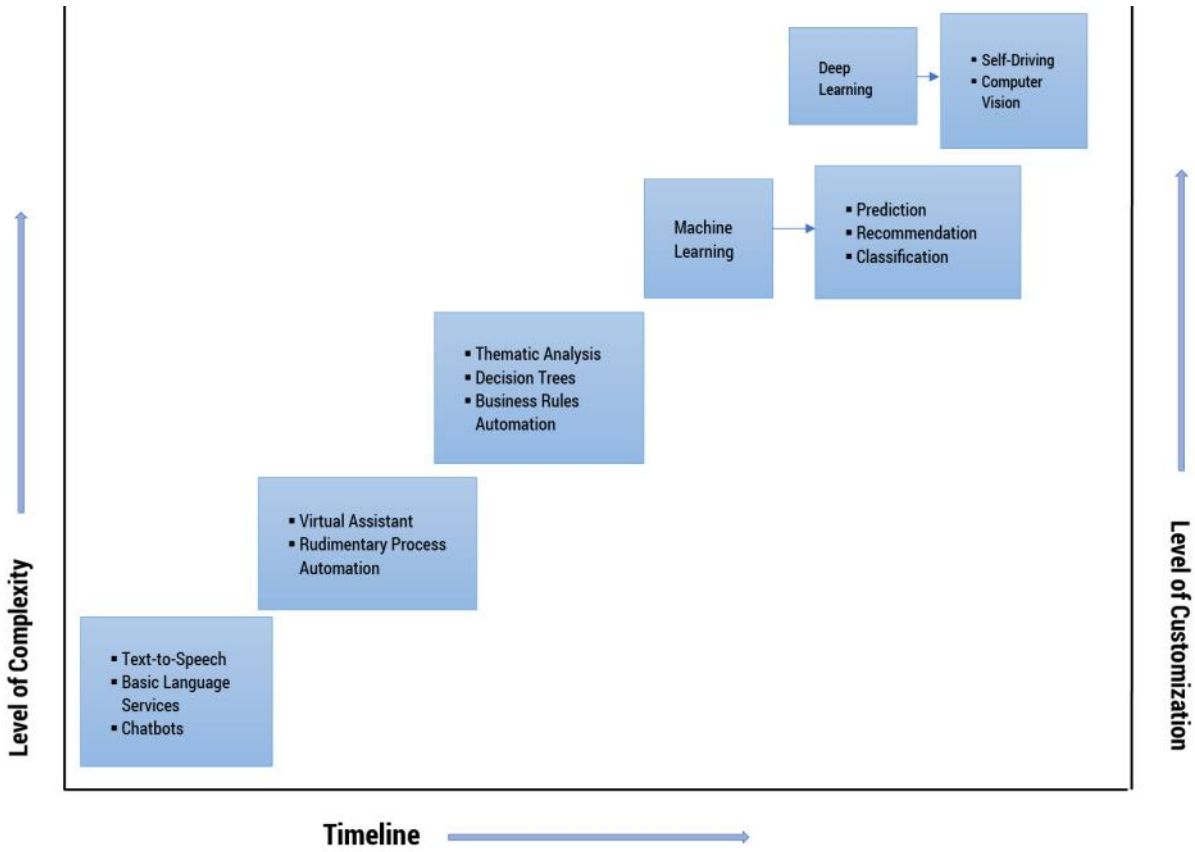

Figure 2 — A gradual approach to AI implementation.

develop more involved use cases with huge business potential. Examples of such involved cases are computer vision and computer-aided diagnosis. Finally, companies can develop new business ideas and business models that rely on AI (self-driving cars is one example).

Some Best Practices to Help Organizations Implement AI Initiatives

Know That Digital Maturity Is Key

An Agile mindset, iterative development cadences, cross-functional collaboration, and metrics-driven measurement are some organizational factors used in assessing digital maturity. Organizations may have to reskill and upskill some employees to achieve the desired process maturity of the company and may need to hire individuals with AI, ML, and digital experience. Digital maturity is developed over a continuum and is not an all-or-nothing exercise. Therefore, all efforts made in developing a digital enterprise help facilitate a smoother AI adoption.

> *Along with data engineers, it is typically the responsibility of product managers and data governance teams to ensure that the AI development team has access to high-quality training data.*

Expect Early Setbacks

Initial failures and setbacks are quite common. However, high rates of failure should not deter organizations from pursuing AI opportunities. Instead, organizations should focus on achieving strategic clarity. They should work toward testing business hypotheses efficiently by failing fast and cheap and avoiding unnecessary spending. Such a lean mindset is crucial, as even failed initiatives help build digital maturity and experience in dealing with the challenges AI presents, which could help companies find eventual success.

Develop Good Training Data

Organizations typically require robust data management for AI initiatives to be successful. The whole process of data collection, cleansing, transformation, and loading is necessary. However, a more subtle challenge is the process of reducing bias in data. Training data is the major determinant of the actions and behaviors of AI products and services. In other words, data in AI applications plays the same role that functional specification does in traditional information systems. The development of training data is a critical step but is also the weakest link in the chain of activities. Most projects fail not because of poor algorithm design or unskilled AI engineers but because of a lack of proper training data sets.

Along with data engineers, it is typically the responsibility of product managers and data governance teams to ensure that the AI development team has access to high-quality training data. It is imperative that subject matter experts and product managers who possess comprehensive understanding of the business domain determine the relevant training data. Their choice should be based on the business rules and must be representative of the underlying data generated by business transactions. Delegating such crucial decisions to implementors may lead to unintended consequences stemming from lack of good-quality training data.

Expect Organization-Wide Impact

AI initiatives require data from multiple sources within the company, meaning that organizations will need to build a knowledge base of corresponding workflows, business processes, and nuances across business domains such as regulations or custom workflows. This effort requires collaboration among several cross-functional units, which can be daunting, especially for traditional companies used to working in functional silos. Traditional companies usually face higher hurdles when middle management (typical power centers within the organization) feels challenged. In such cases, AI initiatives might turn into comprehensive change management and organizational transformation projects.

Recognize That Leadership Is Critical

While AI interventions are vital to an organization's success, they often raise several ethical issues. One such issue is AI's potential for mass job displacement, which is a source of anxiety for operational staff as well. Leaders should reassure their employees about job safety and take steps to reskill their employees. Such leadership will encourage all employees within the company to aggressively leverage AI capabilities in

their day-to-day work to achieve sustained competitive advantage for the organization.

Expect Strict Regulations

Technology enterprises should expect more regulatory hurdles, especially around the privacy of user data. The European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) in the US are already in place; other US states intend to pass rules similar to California's. AI and data-oriented companies may face considerable challenges due to a lack of regulatory clarity as these regulations continue to evolve. Digital organizations need to place special emphasis on cybersecurity, data-specific audits, and regulatory compliance.

Many pioneering companies are exploring federated learning approaches to mitigate these emerging risks. Federated learning is a ML approach where statistical models are executed over remote devices, keeping data local to those remote devices or data centers. Such distribution of data across several remote sources acts as an effective control by preventing access to the whole data set in the event of a security attack.

Conclusion

This article highlights multiple salient facets of AI strategy and adoption planning. A colloquial expression that best captures the current AI market situation is, "The genie is out of the bottle." It is in the best interest of organizations to leverage the "genie" to their advantage by becoming early adopters — but we urge organizations to tread the path with caution. A gradual but consistent approach to AI initiatives should help organizations compete effectively in these early phases of the AI revolution.

References

[1]Hall, Curt. "AI & Machine Learning in the Enterprise, Part I: Current Status." Cutter Consortium Data Analytics & Digital Technologies *Executive Update*, Vol. 18, No. 2, 2018.

[2]Dans, Enrique. "Artificial Intelligence, Buzzwords and Competitive Value." *Forbes*, 25 September 2017.

[3]We reviewed contemporary literature to develop a deeper understanding of state-of-the-art thinking on strategic considerations in AI product development. We also conducted multiple semistructured interviews with seven subject matter experts in the field, including two AI product managers, two founders of an AI company, two AI engineers, and one researcher. We discussed the following three broad topics in interview sessions with each expert: (1) What is an AI strategy? (2) What is the process of developing an AI strategy? and (3) What are the fundamental components of an AI strategy?

[4]Cristianini, Nello. "The Road to Artificial Intelligence: A Case of Data over Theory." *New Scientist*, 26 October 2016.

[5]Gohil, Raveena. "Artificial Intelligence and Machine Learning: How Are They Different?" ACE (blog), 13 August 2019.

[6]"Worldwide Spending on Cognitive and Artificial Intelligence Systems Forecast to Reach $77.6 Billion in 2022, According to New IDC Spending Guide." IDC, 19 September 2018.

[7]Kastrenakes, Jacob. "Microsoft Made a Chatbot That Tweets Like a Tee." The Verge, 23 March 2016.

Pavankumar Mulgund is a Clinical Assistant Professor and doctoral candidate at SUNY Buffalo with more than 12 years' corporate and consulting experience. He extensively researches digital health interventions and their impact on various stakeholders (e.g., patients, providers, payers, and public health organizations), and his interests include technology strategy, the business value of IT, and the application of novel technologies (e.g., AI, IoT, and blockchain) in the context of health information technology. Mr. Mulgund has published several papers in leading academic and industry journals, is a frequent speaker at IS conferences, and has consulted for several organizations. He has developed and taught graduate-level IS courses in database management systems, systems analysis and design, data visualization with Tableau, and experiential IT projects. Before joining SUNY Buffalo, Mr. Mulgund was leading product and delivery teams for a primary contractor of the US Centers for Medicare & Medicaid Services. He has also worked for IBM and Mindtree, among others. Mr. Mulgund holds several professional licenses and expert-level certifications in project management, Agile methodologies, design thinking, and digital business. He can be reached at pmulgund@buffalo.edu.

Sam Marrazzo is CIO of Buffalo Niagara Medical Campus (BNMC). With a focus on driving economic opportunity that will position local companies for future growth, as well as retain and attract young people to the area, Mr. Marrazzo leads BNMC's efforts to develop an innovative, reliable, and secure IT environment throughout the organization and the community. He is responsible for developing innovation strategy and fostering collaboration with industry, education, and startups to help local organizations prepare for, and capitalize on, current and future technologies. Through his role at the BNMC, Mr. Marrazzo helps local companies rethink their business practices and corporate cultures in order to optimize their position, as innovation and technology bring significant changes throughout the economy. With over 25 years in multiple IT roles, his career spans across accounting, supply chain systems, logistics, sales, and transportation, and he has worked at several Fortune *200 companies, including Praxair, Avnet, and McGraw-Hill, in various positions in IT management and strategy, project management, architecture consulting, and development practices. Mr. Marrazzo has been a guest lecturer at the University at Buffalo, and he has been instrumental in attracting and organizing technology-focused events to the area, including the* TopCoder Open *and various hackathons and gaming competitions. Mr. Marrazzo is also a board member of InfoTech WNY. He holds a bachelor of arts degree in business systems from Daemen College, an MBA from SUNY Buffalo School of Management, and served four years in the US Navy. He can be reached at SMarrazzo@bnmc.org.*

THINKING AHEAD

Who Knew THAT Would Happen?

by Paul Clermont

An Unfortunate Coincidence

As artificial intelligence (AI) goes mainstream, putting traditional IT on steroids, it has never been more important for Big Tech to maintain the trust and respect it gained over decades. Well, that long honeymoon has officially ended. There's a new appreciation of how AI capabilities can be used for both good and ill, and the level of public and governmental confidence that they will be used solely for the good is at an all-time low. Indeed, concern and skepticism have emerged as a rare bipartisan position in the US Congress, with Big Tech perceived as having no more of an ethical compass than any other industry, with its protestations to the contrary deemed hypocritical.

The level of public and governmental confidence that AI capabilities will be used solely for the good is at an all-time low.

This happened in great part because of inattention to, and cavalier attitudes about, unintended consequences of innovations introduced at a furious pace. Most of these unintended consequences could and should have been foreseen, but they weren't — at least not by key decision makers. Does this reflect a lack of critical thinking? Probably, because critical thinking likely was not encouraged, and folks asking tough questions went unheard or were sidelined.

With these unintended consequences as our backdrop, the following are the objectives of this article:

- Raise awareness of unintended consequences, especially needed in the context of IT enhanced by AI.
- Discuss how better to identify possible unintended consequences in advance (i.e., preventing the preventable).
- Propose design and management principles for minimizing (bad) unintended consequences.
- Propose public policies that would mandate proper use of algorithms and fix legal responsibility for misuse.

The Law of Unintended Consequences

We're all familiar with Murphy's law: if something can go wrong, it will. Sod's law, perhaps better known in Britain, says it will go wrong in the worst possible way. And then there's the less familiar Finagle's law: it will go wrong at the worst possible time. Such wry humor helps us through inevitable setbacks.

But there's a much more serious law that is nearly universally applicable: the law of unintended consequences. Wikipedia offers a clear and concise articulation: "unintended consequences … are outcomes that are not the ones foreseen and intended by a purposeful action."[1] Wikipedia goes on to group unintended consequences into three types:[2]

1. "*Unexpected benefit* … also referred to as luck, serendipity, or a windfall," which, delightful as they may be, are not the subject of this article.
2. "*Unexpected drawback*: [a] … detriment occurring in addition to the desired effect…."
3. "*Perverse result* … contrary to what was originally intended ([that actually] makes a problem worse), … sometimes referred to as 'backfire.' "

The Rich World of Unintended Consequences

Examples of unintended consequences abound in government and the social sciences, but also occur in engineering, business, management, IT governance, and other areas. Let's look at some examples.

Quotas

Classic sales management tells us that salespeople's productivity is stimulated by commissions and quotas rather than salaries, with supervision required to avoid booking unprofitable sales. But what of customers and potential customers lost by overly aggressive salespeople? (Apple Store salespeople are salaried.)

Measures and Gameable Systems

No Child Left Behind (NCLB) was a well-intended measure to incentivize schools in the US to improve their performance as measured by students' standardized test results, with penalties up to and including closure for schools that did not show adequate improvement. In one school in the state of Georgia, teachers "corrected" students' tests. This case was blatant enough to send a few people to prison, but surely other similar cases have gone undetected. Another unintended but easily foreseeable consequence of NCLB is "teaching to the test," with the presumed objective of broadly educating the students taking a distant second place.

Public Policy

Mandatory minimum sentencing laws, intended to fight crime with draconian punishment, have often created career criminals by sending salvageable young people to prison, giving them records that make them virtually unemployable after release but well schooled in the ways of crime. Where no AI or IT is involved, such laws robotize judges formerly responsible for imposing sentences that reflect not just the crime but also judgment about the offender's character. In some cases, there are automated systems that constrain judges' actions even further, resulting in inequitable and irreversible sentencing.

IT Management

The waterfall methodology for developing applications attempted to bring logic and order to an often chaotic process based on the paradigm that the intended users would describe what they wanted, and the IT people would build it. This approach required gaining agreement on successively more detailed documents describing the end product before serious technical work began, with formal documents and sign-offs thrown over the wall between IT people and the users. The quantity and quality of collaboration to devise approaches that would bring the bulk of the benefit for a fraction of the cost was largely unspecified and usually minimal. The consequence was an annoying bureaucratic process, in practice primarily dedicated to CYA ("cover your ass"), that produced expensive, untimely, and often unsatisfactory results (if any results at all), giving the whole approach such a bad name that it was usually abandoned wholesale — an example of perverse consequences that negated anything good.

IT for the General Public

Numerous examples exist in this area, including:

- Social networks that only a few years ago seemed to empower people against despots have become tools of repression used by despots (e.g., fomenting genocide in Myanmar).
- Democratizing the ability to mass-distribute content worldwide through the Internet so that anyone's voice could be heard provides an unprecedented megaphone for cranks, bigots, mischief makers, and certifiable lunatics to spread not just fake news but bizarre conspiracy theories and poisonous lies.
- Algorithms that help connect people to music and fashions they like are also used to seal them inside informational echo chambers about current affairs, amplifying political divisions.
- The amazingly handy access to online information also provides access to time wasters, and, for large numbers of people, screens seem to be addictive.
- Access to educational material for children also means access to junk programming that turns kids into passive couch potatoes and, worse yet, sells them junk[3] that they can actually order on the spot with their parents' credit cards. The news that many tech executives are strictly limiting their children's access and screen time should be instructive.
- Social networks that help people enrich their connections have, in the hands of adolescents, hugely amplified the voices of the nasty people we all remember from our school years who have harassed their prey even to the point of suicide.

Machines That Haven't Learned Enough

These machines present special risks as AI and AI-enabled robots amplify IT's ability to go wrong, possibly catastrophically, when they fail to properly recognize what they sense, thus failing to react appropriately. Incidents of this include:

- Robots and robotic features in vehicles can kill and maim, as with the recent Boeing 737 MAX crashes.[4]
- Face recognition software, potentially so helpful in ensuring that the guilty and only the guilty are brought to book, has proven highly unreliable with African American faces.[5] In another example, such software identified 28 members of US Congress as matching one of a set of mugshots of known criminals.[6] (Yes, I know, but please.…)
- Medical diagnoses that lead to undertreatment can kill by neglect, while side effects of overtreatment can be devastating.

Rooting out unintended consequences requires both rigor and imagination, with a liberal sprinkling of psychology.

Flawed Algorithms

Flawed algorithms are bad:

- *When* they are found to exhibit bias, whether intentional or not. For example, hiring algorithms that rely on the profiles of the employer's success models may undervalue people, often women, with less traditional career paths.
- *Even* if they're not biased when they lead to walking away from good business or accepting bad business.

Preventing Unintended Consequences: Rounding Up the Usual Suspects

Rooting out unintended consequences requires both rigor and imagination, with a liberal sprinkling of psychology. Human nature and organizational culture can get in the way.

Factors in Human Nature

Face it, we're all flawed to some degree:

- We show confirmation bias when we subconsciously screen out data that doesn't comport with our a priori views.
- Alphas, male and female, who tend to be the loudest voices in the room and, thus the most influential, don't hold back when they should, drowning out tough but needed questions.
- We are not immune to bias however hard we try to fight it.
- We tend to resist change in general, and new, more formal disciplines to root out and deal with unintended consequences are no exception.

Factors in Organizational Culture

Some cultures seem born prudent and others not. The former is much better at identifying and managing unintended consequences:

- Decisiveness is overvalued. Some people, who tend to be leaders, often enjoy making decisions, but are not necessarily better decision makers than others who are more deliberate in their thinking. Decisiveness is generally considered a virtue but being decisive for its own sake can be as dangerous as protracted vacillation. When we don't yet need an absolute decision and there's opportunity to learn more that could affect that decision, why decide at the present time?[7]
- Positiveness is overvalued. "What could go wrong?" is a vital question, and those who are good at asking it and offering possible answers need to be valued as team members, not ostracized as being "negative." Groupthink is the unintended consequence's best friend.
- Daredevil approaches to risk (e.g., "Damn the torpedoes, full speed ahead!"[8]) are occasionally the right tactic in war and business, but not when unintended consequences are likely lurking.
- A "just make your numbers or else" management style, especially when adding, even if only implicitly, "and don't bother me with how you did it" incentivizes failing to seek out, hiding, or ignoring unintended consequences.

- Speed to market is often critical, but shortcuts can be catastrophic (e.g., the Boeing 737 MAX tragedies).
- Clear accountability is critical. There's an old saying that "success has many parents, but failure is an orphan" that applies when success is marred by unintended consequences that could and should have been anticipated.

Envisioning Initiative-Specific Possibilities

Avoiding and mitigating human and cultural issues is necessary to create the right environment to root out and address potential unintended consequences, but that right environment is not sufficient without the imagination to think rigorously about what specific unintended consequences could occur. Typical sources of initiative-specific unintended consequences are:

- Flawed, incomplete, gameable performance measures that can't or won't measure what you care about and may even create perverse incentives for employees.
- Incentives for some people somewhere to do something we'd rather they didn't.
- Unintentionally biased algorithms.
- Machines that haven't yet learned quite enough, as in the cases of self-driving car fatalities,[9] rare as they've been. (There's something particularly awful about an automation killing or injuring a bystander in a way a human-operated machine never would.)
- Inattention to how a well-intended initiative might be perceived by important stakeholders who can kill or hobble it, whatever its merits. (Oh, to see ourselves as others see us.)

Countermeasures

Design Principles

These principles include the following:

- Do no harm. Hippocrates nailed that more than 2,000 years ago, and it's Isaac Asimov's first law of robotics.
- Do some good. Algorithms and machine learning (ML) must be tools that improve human decision making by compensating for our human biases and limited experience.
- Know that because you can do something does not mean you should (e.g., overdesign, as in building connectivity that unnecessarily opens up surface area for intrusion when it's not essential to achieve the primary benefits of the product or service).
- Ensure that the logic of an algorithm is enough to be explainable to responsible authorities in the enterprise using it and be free of surprises (i.e., "features" not revealed before its implementation).
- Establish formal responsibility and custody for algorithms and trained machines that includes initial validation and continuous improvement. The latter includes reverifying effectiveness in their intended use, reassuring freedom from bias based on results, uncovering evidence of gaming, and addressing unintended consequences that slipped past the original design. The organization in which to place this function might best be a compliance department, or at least outside of IT.
- Use algorithms and ML to calculate probabilities. For any loan product, for instance, there is an acceptable default rate. A portfolio that defaults at a higher rate is obviously a problem, but one that defaults at a lower rate represents opportunity loss from rejecting good business. Probabilities in a range around the acceptable default probability should be referred to an experienced loan officer for resolution to avoid arbitrariness based on possibly flawed or ambiguous input.
- Include internal consistency checks in data entered on, for example, loan applications, to help identify anomalies due to errors, gaming, or just the specialness of the case.
- Provide some form of "containment vessel" that can recognize errant functionality and shut it down to keep it from spreading. DeepMind's highly unorthodox but effective move when it beat the *Go* master in 2016 was great in a game,[10] but decisions affecting people's lives are not a game.
- Identify and address external risks, such as data loss, malicious mischief, and hacking into a system to change algorithms and controls.
- Utilize devil's advocates to identify unintended consequences and vulnerabilities and assess their risks.[11] This could be an assigned role on a design team or an independent auditor. IT experts are not necessarily good at this task, so when the stakes are

high enough, experts from disciplines like behavioral economics, psychology, sociology, and ethics — equipped with a cynical edge — can be helpful.

A Caution About Bias

Most people would agree that bias is unacceptable, particularly if it's intentional. Unintentional bias is very hard to recognize in ourselves or in our cultures, so in matters of race, gender, and so on, the term of art for unintended bias is "adverse impact" on some defined and usually historically disadvantaged group. In practical fact, bias is almost impossible to eliminate 100% in any nontrivial algorithm. That said, it's incumbent on the designer of an algorithm to make all reasonable efforts to assure the minimization of bias and on the user to ensure that the results so reflect.

> *Transparency is a good thing, but like a lot of good things, it can be overdone, at least with respect to algorithms.*

A Closer Look at Transparency

Transparency is a good thing, but like a lot of good things, it can be overdone, at least with respect to algorithms. "Need to know" is critical. Clearly, an enterprise that uses an algorithm designed and built by others must understand the logic of that algorithm, since the enterprise is responsible for the algorithm's results. However, front-line people who collect data fed into the algorithm do not need to know and should not know (or be too easily able to deduce) specific details, lest they game the input to incorporate prejudices or personal feelings.

Public Policy

Some practices in the formulation and use of algorithms may need the force of laws and regulations to become standard:

- For algorithms, an analogue to the Underwriters Laboratory that certifies safety of electrical gizmos could be established to certify freedom from bias.
- There must be a human-based review of borderline results of life-affecting algorithms (e.g., home mortgages, employment, criminal justice, and medical diagnoses) with a requirement to document algorithm-produced decisions and human interventions.
- No black boxes! Algorithm vendors must explain their algorithms' underlying logic to their customers' responsible staff — the "custodians" cited above — whose formal acceptance fixes responsibility for the algorithms' results, absent a failure of the vendor to disclose some hidden feature or fix a bug. Nondisclosure agreements can cover proprietary aspects, but the user must not be allowed to plead ignorance of what was in the black box as a defense against accusations of bias or incompetence.
- Unlike an algorithm, we cannot easily explain pattern recognition software; its input is a picture (or text) rather than digital data entered by a human. Purchase by a medical practice, for example, of software that has learned how to recognize malignant skin lesions should not relieve clinicians of responsibility for diagnoses.
- Legal responsibility for unintended consequences, regardless of tests by anyone, should fall under the enterprise using the algorithm.
- The ultimate test for bias lies in the results produced by the algorithm, not in its logic or its designers' intentions.[12]

A Matter of Urgency

When technology is good, it is very, very good. It makes us more productive and mobile, reduces drudgery, opens up new ways to enjoy life, makes us more safe and secure. In other words, it's how we came to live in the world of today rather than in caves, hunting and gathering. But when technology is bad, and when a major form of it — IT — goes on AI-based steroids, the opportunity for bad increases dramatically. While no credible argument has been made that familiar players in the industry have acted with malevolent intent, undesirable results have nonetheless manifested themselves and, unfortunately, some significant industry leaders have come across as tone-deaf and cavalier in their responses. Hence, the last few years' loss of innocence among the public and their elected representatives.

The time is now for Big (and Small) Tech to focus on regaining lost trust and credibility. That won't be easy. It means publicly and straightforwardly accepting

responsibility for past missteps, without minimizing or dissembling (or appearing to), and maintaining that approach to future mishaps.[13] This is a job for CEOs, not public relations flaks. CEOs need to be clear and specific about what they intend to do, including new approaches befitting the seriousness of the problems that range from simply translating privacy-related terms and conditions into clear and concise language all the way to fundamentally modifying or even shutting down problematic products and services.

Now is also the time for technology companies to proactively reach out to legislators and regulators around the world with appropriate openness and humility, showing by words and deeds how seriously they take basic issues and unintended consequences. The goal is cooperative and creative problem identification and solution, without explicit preconditions or no-go areas. Not every government entity will necessarily be receptive, but enough probably will be to promote the recovery of trust and credibility.

That is Plan A. There is no Plan B that is at all pleasant to think about. For tech to proceed as if nothing had really changed would be a disaster, exposing the industry to endless lawsuits and overly reactive (and probably ham-handed) regulation. That scenario is also not good for society. We will always need the fruits of creative inventors and entrepreneurs who bring those fruits to market, but we need them to do so responsibly, thinking through and mitigating unintended consequences much better than they have thus far. Brilliance and daring need to be tempered with broader perspectives on what may happen in the real world.

Will the pace of innovation slow down? Yes. Is that bad? No, as we've learned. Thinking ahead and seeing around corners have always been hallmarks of prudent management and sustained success and always will be.

References

[1]"Unintended consequences." Wikipedia.

[2]Wikipedia (see 1).

[3]This point echoes critics from the early years of television. In 1961, US Federal Communications Commission (FCC) Chair Newton Minow characterized TV as a "vast wasteland," a quote that has never stopped being repeated — or relevant.

[4]"Boeing 737 MAX groundings." Wikipedia.

[5]Simonite, Tom. "The Best Algorithms Struggle to Recognize Black Faces Equally." *Wired*, 22 July 2019.

[6]Brandom, Russell. "Amazon's Facial Recognition Matched 28 Members of Congress to Criminal Mugshots." The Verge, 26 July 2018.

[7]This seemingly countercultural view was espoused by no less than Robert Rubin, former US Secretary of the Treasury and CEO of Goldman Sachs; see: Weisberg, Jacob. "Keeping the Boom from Busting." *The New York Times Magazine*, 19 July 1998.

[8]Attributed to Admiral David Farragut during the Battle of Mobile Bay in the American Civil War. Fortunately for him, his sailors, and his reputation, none of those "damn torpedoes" hit their targets.

[9]"List of self-driving car fatalities." Wikipedia.

[10]Moyer, Christopher. "How Google's AlphaGo Beat a Go World Champion." *The Atlantic*, 28 March 2016.

[11]In a recent Massachusetts Institute of Technology (MIT) conference on the future of IT, a speaker from Salesforce said they had instituted such a discipline.

[12]A recent proposal from the US Department of Housing & Urban Development (HUD) would make the burden of proof of adverse impact much higher; see: Capps, Kriston. "How HUD Could Dismantle a Pillar of Civil Rights Law." CityLab, 16 August 2019.

[13]This general rule for doing business is honored mostly in the breach, leading to the cover-ups that end with consequences worse than those of the offense they're covering up. (Anyone remember Watergate?)

Paul Clermont is a Senior Consultant with Cutter Consortium's Business Technology & Digital Transformation Strategies practice. He has been a consultant in IT strategy, governance, and management for 30 years. His clients have been primarily in the financial and manufacturing industries, as well as the US government. Mr. Clermont's major practice areas include directing, managing, and organizing IT; reengineering business processes to take full advantage of technology; and developing economic models and business plans. He is known for successfully communicating IT issues to general managers in a comprehensible, jargon-free way that frames decisions and describes their consequences in business terms. In his consulting engagements, he follows a pragmatic approach to the specific situation and players at hand and is not wedded to particular models, methodologies, or textbook solutions. Mr. Clermont has spoken and written about the challenges of getting significant and predictable value from IT investments and has taught executive MBA courses on the topic. His undergraduate and graduate education at MIT's Sloan School of Management was heavily oriented toward operations research. He can be reached at pclermont@cutter.com.

CUTTER Business Technology Journal

About Cutter Consortium

Cutter Consortium is a unique, global business technology advisory firm dedicated to helping organizations leverage emerging technologies and the latest business management thinking to achieve competitive advantage and mission success. Through its research, training, executive education, and consulting, Cutter Consortium enables digital transformation.

Cutter Consortium helps clients address the spectrum of challenges technology change brings — from disruption of business models and the sustainable innovation, change management, and leadership a new order demands, to the creation, implementation, and optimization of software and systems that power newly holistic enterprise and business unit strategies.

Cutter Consortium pushes the thinking in the field by fostering debate and collaboration among its global community of thought leaders. Coupled with its famously objective "no ties to vendors" policy, Cutter Consortium's *Access to the Experts* approach delivers cutting-edge, objective information and innovative solutions to its clients worldwide.

For more information, visit www.cutter.com or call us at +1 781 648 8700.