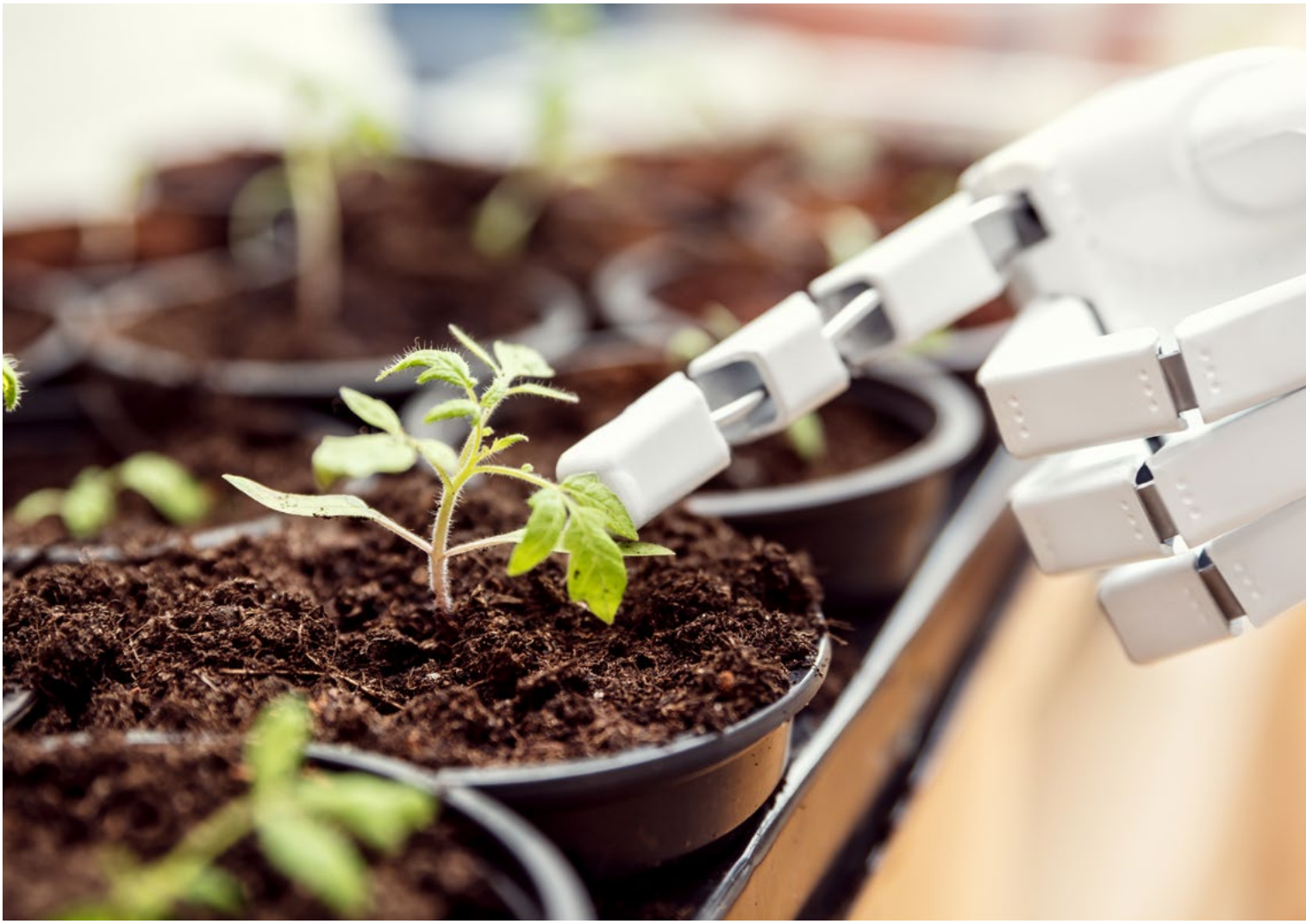
protiviti®

Face the Future with Confidence



Artificial Intelligence: Can Humans Drive Ethical AI?

Building AI systems with ethical principles and values

Internal Audit, Risk, Business & Technology Consulting

Introduction

Artificial intelligence (AI) is a powerful technology that's driving innovation, boosting performance, and improving decision-making and risk management across enterprises. It's also turning data into the key driver of competitive advantage. Over the next two years, organizations across all industries plan to deploy or increase their use of artificial intelligence, according to a recent global executive survey on AI conducted by Protiviti.[1] AI will be such a significant game-changer that those who don't take full advantage of the technology will find themselves at a competitive disadvantage.

Businesses using advanced AI today are already seeing value in the form of reduced costs, accelerated time-to-market, increased customer retention and improved employee engagement. But AI also can introduce potential ethical and social consequences. As AI-enabled technologies become an ingrained part of our everyday business world, it is important to step back and look at the societal and moral implications.

"In the near future, many of us will find ourselves working with or alongside AI-enabled technology. It's up to humans to ensure that AI systems are designed with the right algorithms, fed the right data and monitored properly to keep them aligned with the best goals and values of humanity."

- Ron Lefferts, Managing Director and Leader of Protiviti's Global Technology Consulting Practice

[1] *Competing in the Cognitive Age*, Protiviti, 2018: www.protiviti.com/sites/default/files/united_states/insights/ai-ml-global-study-protiviti.pdf.

What Is "Ethical AI"?

Should AI systems make decisions about human health and safety? Can we trust a robot's analysis and efficiency in areas where compassion or justice matter? Like an airplane, which can accelerate travel or drop bombs, or the internet, which can keep us informed but also spread misinformation, AI is simply a tool. It's up to people to guide what AI *should* do based on our best societal norms and human morality. In practical terms, that means we must take a mindful and active role in determining how AI algorithms are designed and what data is used to train the machines. We must also take charge of monitoring and correcting the behaviors that AI systems are learning from the data over time.

To formalize its expectations of AI, the European Commission issued AI guidelines[2] as part of its Digital Single Market initiative. According to the guidelines, AI systems must be lawful, ethical and robust in order to be trustworthy. In other words, AI systems must respect all applicable laws and regulations, as well as ethical principles and values. And AI systems must also function robustly, not only from a technical perspective, but considering the full social environment.

[2] "Ethics guidelines for trustworthy AI," *Digital Single Market*, April 8, 2019: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

To meet this standard, the guidelines include the following key requirements for AI:

01 **Human agency and oversight —** AI systems should be designed with human-in-the-loop, human-on-the-loop, and human-in-command approaches so that humans always have the last word on the system's actions.

02 **Technical robustness and safety —** AI systems should be resilient, secure, accurate and reliable to minimize and prevent harm.

03 **Privacy and data governance —** AI systems must respect privacy and ensure protection of data. Data privacy controls should be part of the design and should encompass data quality, data integrity and secure data access.

04 **Transparency —** AI systems, data and business models should be explainable in a way that human stakeholders understand. Humans must be made aware when they're interacting with AI systems, and must be informed about the systems' capabilities and limitations.

05 **Diversity, nondiscrimination and fairness —** AI systems should avoid bias and be accessible to everyone, regardless of ability. Relevant stakeholders should be involved throughout the systems' life cycle.

06 **Societal and environmental well-being —** AI systems should benefit all human beings, including future generations, and consider environmental responsibility and societal impacts.

07 **Accountability —** AI systems should include mechanisms to ensure accountability and adequate access to redress. Algorithms, data and design processes should all be auditable.

National Strategies for Ethical AI

More than 15 countries have released national strategies to promote the use and development of AI, including ethical standards. Several organizations and countries have also taken interest in regulating AI. Some of the most recent initiatives include the following:

- In 2016, the Obama administration produced two reports that set priorities, enlisted key individuals, and produced a comprehensive national plan for AI.[3] In February 2019, President Trump signed an executive order to spur investment in AI. However, Brian Tse, with the Center for the Governance of AI at the University of Oxford, noted the order showed "insufficient discussion on ensuring that AI systems can be developed and used within an ethical and responsible framework."[4]
- In 2017, Canada and China each released an AI strategy. China's plan is the most comprehensive of any country's plan, and includes initiatives and goals for regulations and ethical norms.[5]
- In February 2019, the Organisation for Economic Co-operation and Development (OECD) Expert Group on Artificial Intelligence in Society (AIGO) announced it will launch an AI Policy Observatory,[6] whose purpose is to be a "center for evidence collection, debate and guidance for governments on how to ensure the beneficial use of AI." OECD has 36 member nations.
- In April 2019, the Institute of Electrical and Electronics Engineers' (IEEE's) Global Initiative on the Ethics of Automated and Intelligent Systems published its first edition of "Ethically Aligned Design," a guide for governments, businesses and the general public that is aligned with society's defined values and ethical principles.[7] IEEE is the world's largest professional association dedicated to advancing technology, with approximately 140 nations represented.

[3] Agrawao, Ajay, et al. "The Obama Administration's Roadmap for AI Policy." *Harvard Business Review*, 21 Sept. 2017, hbr.org/2016/12/the-obama-administrations-roadmap-for-ai-policy.

[4] "Oxford University AI Policy Researcher Says Trump's AI Initiative Falls Short on Immigration and Ethics Issues." *Synced*, 22 Feb. 2019, syncedreview.com/2019/02/22/oxford-university-ai-policy-researcher-says-trumps-ai-initiative-falls-short-on-immigration-and-ethics-issues/.

[5] Dutton, Tim. "An Overview of National AI Strategies." Medium, Politics + AI, 28 June 2018, medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

[6] "Artificial Intelligence — Organisation for Economic Co-Operation and Development." *OECD*, www.oecd.org/going-digital/ai/oecd-initiatives-on-ai.htm.

[7] "SA — The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems." *IEEE*, standards.ieee.org/industry-connections/ec/autonomous-systems.html.

Meeting the Ethical Standard

From a technical perspective, there are several factors that can either bring AI closer to its ethical goal or cause it to deviate from it. We discuss these factors below.

The Importance of Proper Training

When the AI application is under development, data is used to train and test the algorithm. The data fed into the AI algorithm influences the outcome in the same way that experience influences humans. The algorithm detects patterns within the data and uses those patterns to determine its actions. In production, the AI application is consuming data, acting on it, and "learning" from it. As the system is exposed to more and more data, it gets better at learning and is able to optimize the algorithm to achieve better performance, which can lead to new insights and better decision rules. Because an AI system is capable of a processing speed and capacity far beyond those of humans, it can learn and begin to develop its own decisions rules quickly, often moving in an unanticipated direction. For this reason, human supervision must be an integral part of AI applications.

The two main areas where AI could go astray are biases that originate from algorithm design and biases that creep in via the data. One dramatic example of how algorithm bias could affect AI performance is the self-driving car. Let's say the car's AI system is programmed to respond to unavoidable collisions by minimizing loss of life at all costs. That might mean an algorithm scenario in which the car's occupants don't survive, if that choice would save more pedestrians and passengers in other vehicles. Alternatively, the AI system may be programmed to protect the passengers of the vehicle at all costs, meaning pedestrians and passengers in other vehicles are less likely to be saved. Algorithm design, therefore, must take into consideration the context in which it will operate and must balance the needs for efficiency and profitability with the best values of society — not an easy task.

Data Bias

Data bias is a well-known problem to AI researchers and developers. There are many examples of AI-enabled processes that have been marred by data bias, in recruiting, loan application approvals, facial recognition software and more. If loan officers have historically made biased decisions in rejecting individuals belonging to a certain race, gender or age group, for example, the development data for the AI will reflect these biases. If an AI/machine learning (ML) model is developed on this data to predict the risk of loan defaults, it will most likely discriminate against extending credit to individuals belonging to these groups. The scarcity of comprehensive, unbiased or current data sets for training algorithms contributes to the data bias problem — and if everyone uses the same training sets, the bias perpetuates throughout AI applications.

"Data is the most important element in AI, but we have amassed more data than any human could analyze or interpret alone. Using AI, we can analyze large volumes of data to find patterns or solve problems, realizing new connections and yielding greater insights to guide decision-making."

— Madhumita Bhattacharyya, Managing Director, Protiviti Enterprise Data and Analytics

To demonstrate the significance of data bias, MIT scientists created an AI-enabled image-captioning system named Norman. The scientists trained Norman on visual data culled from the internet that contained mostly gruesome or disturbing images. After consuming this data, Norman was tested on the

interpretation of Rorschach inkblots side-by-side with another image-captioning system that had been trained on "healthier" images. Norman saw the inkblots largely as the byproduct of murder or violence, while the healthy system interpreted them as flowers, people, or other innocuous objects. The experiment illustrates the crucial importance of data input to the conclusions and decision-making of AI.

Validation and Testing

To reduce the risk of algorithm bias at the outset, or data bias creeping in over time, organizations should validate and monitor the performance of their AI systems.[8] Companies must also ensure that the data used to train the system is of good quality, and that the data sets used are large enough to ensure variety. The greater the quality, quantity and variety of data used to train the AI system, the better the system will be. Protiviti's AI survey results note that there are four key areas that need to be addressed to ensure the validity of an AI system's performance:[9]

- **Conceptual soundness.** Analyze the AI model's design, review its documentation, assess empirical evidence, and verify that the variable selection process is sound and unbiased.
- **Process verification.** Make sure the model is subject to a validation and approval process. Depending on the system, this could take various forms. For instance, if a model dynamically redesigns its algorithm, the company should have the capability to save all versions of a model to test the validity of changes.
- **Ongoing monitoring.** Monitor model risks and limitations regularly, particularly when systems use automated processes to redevelop the models on their own. Validation teams should have the capabilities needed to monitor these programs on an ongoing basis.
- **Outcomes analysis.** Monitor model performance over time. Simpler models can lead to higher levels of bias, while more complex models can lead to greater variance.[10]

Societal Considerations — Jobs

AI has the potential to deliver efficiencies, decrease errors and create wealth quickly. As companies begin to implement AI and machine learning to automate repetitive tasks in the workplace, employees whose jobs are replaced by technology-enabled automation may fear becoming unemployed. But no matter how effective AI may be, it will not replace humans. Instead, people and AI-enabled technology will work together, because people will have to review every output of an AI or ML model. Instead of eliminating jobs, AI will enhance and transform jobs, creating opportunities for job enrichment and new experiences and challenges. Society faced a similar challenge during the Industrial Age, when assembly lines were introduced and machines began to do the work that was previously done by human hands. Even there, machines did not replace human jobs; instead they transformed and elevated them.

AI will eliminate some jobs, but it will create even more jobs than it eliminates. Employers will need to lead the charge to retrain workers to take on jobs that are more challenging, interesting and meaningful. Because stewardship of AI must always remain the work of human beings, the workforce capacity liberated by AI could be redirected to this richer set of responsibilities, including analyzing AI outputs and monitoring AI systems as they learn and progress.

[8] "Validation of Machine Learning Models: Challenges and Alternatives." Protiviti, www.protiviti.com/US-en/insights/validation-machine-learning-models-challenges-and-alternatives.

[9] ibid, p. 1.

[10] Variance is the degree to which the model can deviate from the mean value.

Goals, Governance and the Organization

An organization's AI strategy should be set with the active engagement of the CEO and board of directors, who should be able to articulate clear goals with respect to the AI program, as well as clear ethical standards that should guide it. The goals set by senior leaders must ensure AI implementations are aligned to desired business outcomes and include human agency and oversight, privacy and data governance, transparency, fairness, sustainability and accountability. These high-level goals must guide the algorithm design and set the standard for initial testing and continuous monitoring. To keep AI applications aligned to business outcomes, AI initiatives should have executive buy-in and be led by line-of-business leaders.

The organizational structure must ensure and facilitate close collaboration between AI experts and business partners. Because AI encompasses both technology adoption and changes to business processes, an AI "center of excellence" should work in coordination with additional AI expertise located within business units. An organizational structure that enables AI specialists and business partners to collaborate provides the best outcome. With too much technical emphasis, AI applications could fail to deliver the intended business benefits. With too much business emphasis, AI applications may not be technically sound or may begin to deviate from the preset standards.

Lucas Lau, Protiviti's director of machine learning/deep learning, observes that the three skill sets critical to AI — business knowledge, data science and data engineering — are rarely found in one individual, but that both AI teams and leadership need access to these skills. The most successful organizations are likely to develop these AI skill sets internally. For others, partnering with universities and competent third parties or outsourcing AI development may be a more feasible course. In either case, knowledge transfer should be a component of any AI engagement or initiative.

Conclusion

Although an ominous future where humans are ruled by machines is a popular trope in films, machines have no inherent malice. AI systems are neutral — but the old adage of "garbage in, garbage out" has never been more apt.

"In the near future, many of us will find ourselves working with or alongside AI-enabled technology," says Ron Lefferts, managing director and leader of Protiviti's global Technology Consulting practice.

Protiviti's AI survey results predict an explosion in AI business applications in the next two years. These intelligent machines will not replace human beings, but they could dramatically alter the way we work and live. Lefferts adds, "Starting with clear goals, it's up to humans to ensure that AI systems are designed with the right algorithms, fed the right data and monitored properly to keep them aligned with the best goals and values of humanity."

How Protiviti Can Help

Protiviti's interdisciplinary teams help solve our clients' unique business challenges using data and analytics and leveraging technologies such as AI and machine learning. Our professionals bring deep industry expertise and extensive technology and consulting experience to implement technical solutions and change programs that enable clients to create a competitive advantage and capitalize on financial benefits from adopting AI/ML. We leverage our deep domain expertise and advanced analytical horsepower to enable sustainable, business-relevant, technology-enabled and globally managed operational changes.

ABOUT PROTIVITI

Protiviti is a global consulting firm that delivers deep expertise, objective insights, a tailored approach and unparalleled collaboration to help leaders confidently face the future. Protiviti and our independently owned Member Firms provide consulting solutions in finance, technology, operations, data, analytics, governance, risk and internal audit to our clients through our network of more than 75 offices in over 20 countries.

We have served more than 60 percent of *Fortune* 1000® and 35 percent of *Fortune* Global 500® companies. We also work with smaller, growing companies, including those looking to go public, as well as with government agencies. Protiviti is a wholly owned subsidiary of Robert Half (NYSE: RHI). Founded in 1948, Robert Half is a member of the S&P 500 index.

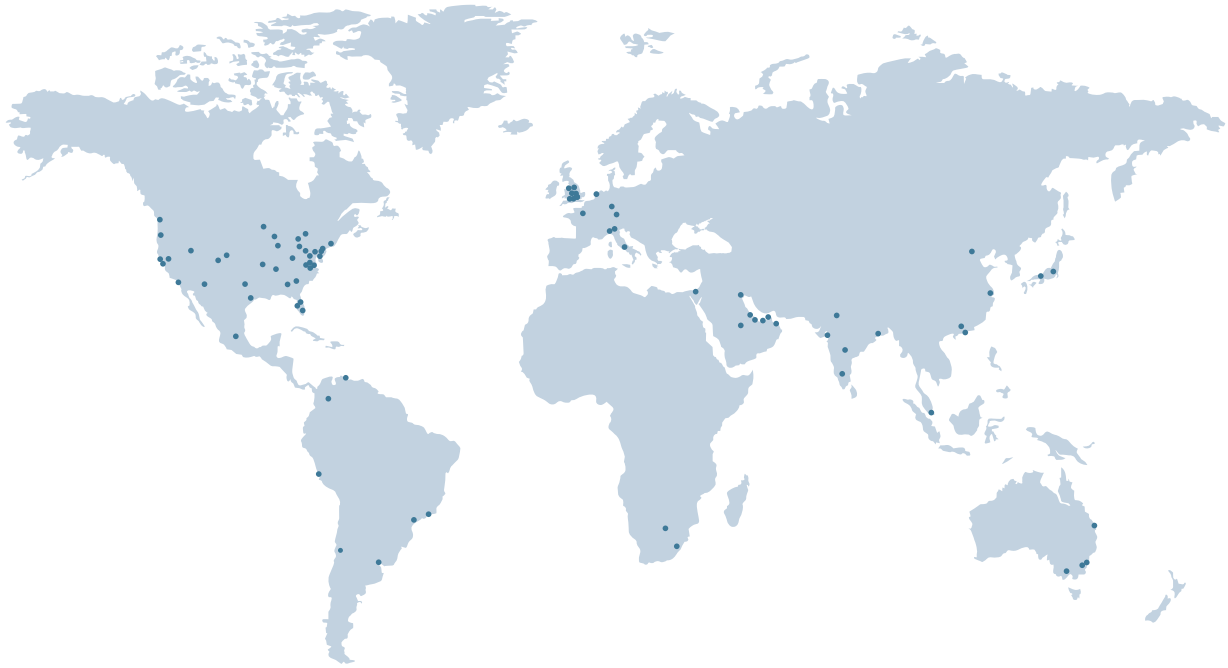
CONTACTS

Madhumita Bhattacharyya
Managing Director, Protiviti Enterprise Data and Analytics
+1.469.540.2119
madhumita.bhattacharyya@protiviti.com

Ron Lefferts
Managing Director, Global Leader of Protiviti Technology Consulting
+1.212.603.8317
ron.lefferts@protiviti.com

ACKNOWLEDGEMENT

Contributors to this white paper include Shaheen Dil and Lucas Lau.



THE AMERICAS

UNITED STATES
Alexandria
Atlanta
Baltimore
Boston
Charlotte
Chicago
Cincinnati
Cleveland
Dallas
Denver
Fort Lauderdale
Houston
Kansas City
Los Angeles
Milwaukee
Minneapolis
New York
Orlando
Philadelphia
Phoenix
Pittsburgh
Portland
Richmond
Sacramento
Salt Lake City
San Francisco
San Jose
Seattle
Stamford
St. Louis
Tampa
Washington, D.C.
Winchester
Woodbridge

ARGENTINA*
Buenos Aires

BRAZIL*
Rio de Janeiro
Sao Paulo

CANADA
Kitchener-Waterloo
Toronto

CHILE*
Santiago

COLOMBIA*
Bogota

MEXICO*
Mexico City

PERU*
Lima

VENEZUELA*
Caracas

EUROPE, MIDDLE EAST & AFRICA

FRANCE
Paris

GERMANY
Frankfurt
Munich

ITALY
Milan
Rome
Turin

NETHERLANDS
Amsterdam

UNITED KINGDOM
Birmingham
Bristol
Leeds
London
Manchester
Milton Keynes
Swindon

BAHRAIN*
Manama

KUWAIT*
Kuwait City

OMAN*
Muscat

QATAR*
Doha

SAUDI ARABIA*
Riyadh

UNITED ARAB EMIRATES*
Abu Dhabi
Dubai

EGYPT*
Cairo

SOUTH AFRICA *
Durban
Johannesburg

ASIA-PACIFIC

AUSTRALIA
Brisbane
Canberra
Melbourne
Sydney

CHINA
Beijing
Hong Kong
Shanghai
Shenzhen

INDIA*
Bengaluru
Hyderabad
Kolkata
Mumbai
New Delhi

JAPAN
Osaka
Tokyo

SINGAPORE
Singapore

*MEMBER FIRM

protiviti®