Deloitte.

Artificial Intelligence and rulemaking

Summary of Deloitte publications and policy implications

May 2023

Re these

Contents

Introduction

Artificial Intelligence ("AI") can change the world for good in unimaginable ways. However, AI comes with risks too. Due to the scale and speed with which AI can impact society, thoughtful policy and regulation is needed to mitigate risk.

Overview and key themes

The Deloitte network has significant experience in conceptualizing, developing, and implementing new technologies that support government, commercial, and societal advancement. Whether working with commercial or government clients, this experience has resulted in a proliferation of specialist advice and thought leadership on AI — its promise, but also its limitations. In 2018, Deloitte Global put forward a "better" regulation agenda concept that considered the ideal attributes for regulating technologies such as AI. This regulatory framework proposed methods with the aim to enhance and enable the most positive attributes of new technologies, while protecting individuals, businesses and societies from its potentially negative implications.

In addition, Deloitte as an organization has published extensively on AI. This paper provides a high-level summary of key AI-related papers written to date, with a focus on the policy implications of each paper. For readers interested in each paper's details, the "Review of papers" section includes a paper-by-paper summary and analysis. Additional research could shed light on further policy implications and best practices.

Three consistent themes recur across all papers. These themes speak to the heart of how policy should be created to mitigate the risks of AI but, at the same time, balance these concerns with the opportunities and benefits that can come from AI through innovating and development of products, services, and interaction with AI.

Bias, ethics, and fairness

AI can play an integral role in advancing (or hindering) fairness across society. When used appropriately, AI can help identify and mitigate bias, leading to more fair outcomes for individuals and groups. However, because AI is built and directed by humans, those designing, monitoring, utilizing, and regulating AI should take a proactive approach to identifying and minimizing bias in all its forms, so that AI does not perpetuate existing inequities or create new ones.

Policymakers are especially concerned about the potential negative implications for bias and fairness. Debates over how to regulate algorithms, assess for biased inputs or outputs, and best use AI to advance fairness are active across Europe and the United States. Proposals are the most advanced, and the most far-reaching, in Europe.

For the most part, to date, this issue has been left to the private sector to develop its own best practices to mitigate bias. Deloitte eminence focuses heavily on the importance of understanding the many forms that AI-related bias can take, and how to mitigate such bias across an organization – in line with the view of policymakers and regulators around the world.

Trust

A second theme throughout Deloitte's publications is the role of trust in integrating AI – and the potential it can bring to society – into individuals' daily lives. Trustworthy AI is ethical, lawful, and technically robust; understanding *how* and *why* AI comes to the conclusions that it does is critical to gaining broad trust in the technology. As a result, there are policy implications related to enhancing algorithm transparency and accountability, that is, helping individuals better understand the technology, its processes, and its conclusions.

Much of the policy conversation around trusted AI relates to organizational policies, not regulatory requirements. However, the same challenges and debates present themselves when considering at a national level how to build trust in the use of AI within an economy.

Deloitte US developed the Trustworthy AI framework as a critical response to this; more detail on this framework is in the next section.

Innovation and development

The role of AI in furthering economic prosperity for the societies that harness it is a third theme of Deloitte's eminence. Implemented well, AI could automate and accelerate certain tasks, freeing up workers for higher-value (and higher paid) skilled work. To realize AI's promise on this front, though, it is critical to continue to invest in innovation and the development of AI-related technologies. Whether this occurs at a national level or inside organizations themselves, it is necessary to apply both resources and specialist experience across the public and private sectors to continue development of this technology.

Particularly in the United States, policymakers view AI as a key to enhancing US economic security and competitiveness – as well as national security – into the coming decades. As a result, there are numerous proposals to invest heavily in the development of AI and related research, with several proposals considering government-led investments or public-private research partnerships. Similarly, the EU sees AI as a promising technology that could contribute to solving some of the world's biggest problems in health, the environment, education, and mobility. As a result, the EU aims to create broad enabling conditions for AI technologies to succeed in the EU through acquiring, pooling and sharing policy insights, tapping into the potential of data, and fostering critical computing infrastructure. Additionally, the EU is working on the , creating the first international regulation on AI. In Australia, governments and science organizations have co-developed various papers to discuss what a national AI ethics framework could look like, to respond to issues associated with AI and ensuring ethical and inclusive values are used to manage the deep influence it will have on the way people live, work and play.

Trustworthy AI^{TM}

Deloitte’s Trustworthy AI framework provides a vision for how companies and governments could be responding to the challenges of AI.

Deloitte’s Trustworthy AI framework brings together each of the themes discussed in the introduction.

The framework is the result of interviews with subject matter specialists, data and computer scientists, mathematicians, ethicists, and others. It posits that AI has the potential capabilities to transform economies, workplaces, and individuals’ lives. However, to achieve AI’s promise, all of us – governments, businesses, consumers, and individuals – must trust that its outputs are as intended or, in the least, satisfy society’s basic expectations. The framework proposes six, interrelated dimensions for AI to broadly earn trust. Incorporating these criteria could advance trustworthy AI.

Deloitte’s State of AI in the Enterprise, 5th Edition Study of Enterprise AI Adopters found that 50% of respondents identified AI-related risks as a barrier to scaling their AI initiatives.

Furthermore, 79% of organizations were not educating workers to use AI effectively in their roles.

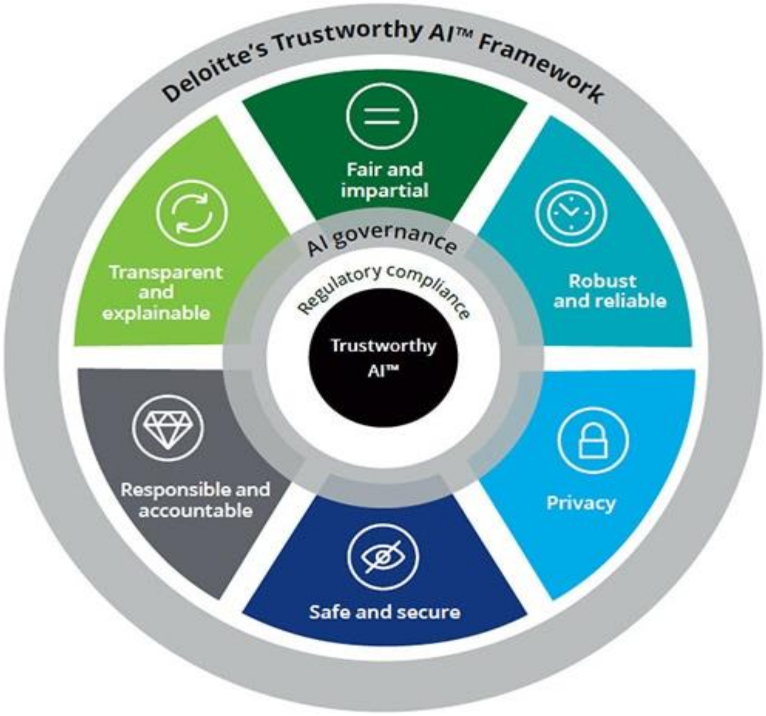


Figure 1, Deloitte’s Trustworthy AI Framework wheel, Deloitte US.

The trustworthy AI framework

Fair and impartial

Assess whether AI systems include internal and external checks to help enable equitable application across all participants.

The definition of "fair" is a challenging one, and organizations should determine what they mean by "fair" before they can train their algorithms to *be* fair. They should also actively search for bias, making adjustments to algorithms and data as necessary.

Transparent and explainable

Help participants and stakeholders understand how their data can be used and how AI systems make decisions. Algorithms, attributes, and correlations are open to inspection.

Participants should understand how their data is collected and used and how algorithms make decisions. This pressure exists for a range of current organizations: online retailers who use data to provide individualized product recommendations as well as legal systems that use AI to inform sentencing decisions. The authors also note a growing pressure to proactively inform individuals when they are interacting with AI.

Responsible and accountable

Put into place an organizational structure and policies that can help clearly determine who is responsible for the output of AI system decisions.

This includes identifying who is responsible and accountable for AI outputs. If an AI causes a problem, is the programmer responsible? The individual interacting with the AI? The CEO? Identifying these responsible parties is also relevant in addressing any problems or negative impacts caused by an errant AI decision.

Robust and reliable

Confirm that AI systems can learn from humans and other systems and produce consistent and reliable outputs.

AI should be as robust and reliable as the traditional systems and processes it is designed to improve or replace. It must consistently produce reliable outputs (e.g., a health care company using AI to identify abnormalities in a scan), even when provided with new data sets. The "human factor" is thus critical.

Respectful of privacy

Respect privacy and avoid using AI to leverage customer data beyond intended use.

AI should comply with data regulations and use data only for stated and agreed-upon uses. Consumers should also have transparency over the way their data is used, and have control over that data, including the right to opt in, or out, of data sharing.

Safe and secure

Protect AI systems from potential risks (including cyber risks) that may cause physical and digital harm to users.

Without protection from cybersecurity risks (internal and external), there could be financial, reputational, or even physical harm (e.g., hacking an AI-enabled finance system, an autonomous vehicle, or the breach of sensitive personal information like biometrics).

Better regulation

In 2018, Deloitte US published perspectives on the future of regulation to provide a foundation for policymakers to consider how legal frameworks can and should keep pace with rapid technological change.

The advancement of transformational technologies like AI are challenging and uprooting outdated systems and practices, replacing them with new models better suited for tomorrow. To unleash their full potential, and appreciate their complexities, regulation and approach to rulemaking should keep pace.

Failure to advance nuanced, risk-based regulation can have its own costs: failing to ensure key privacy or safety protections, hindering innovation, or delaying implementation of transformational technologies, among other outcomes. This is not a matter of more, or less regulation, but of facilitating *improved* regulatory outcomes that serve the public interest.

A "better" regulatory agenda anticipates all the variables of disruptive technologies and their impacts – speed and scalability for governments, business, innovators, civil society, and others – and works together with impacted stakeholders to find balance and prepare for the future.

In short, a better regulatory agenda highlights three core pillars:

- **Targeted conception:** rulemaking that is clear in scope, purpose, and administration.
- **Smart design:** rulemaking that is flexible, innovative, and complementary.
- **Committed implementation:** rulemaking that is enforceable, has proper oversight, and is accountable.

These pillars, in turn, are complemented by key principles driving "better" regulatory practices: transparency, agility, outcome-focused guidelines, grounding in data and evidence, collaborative approaches, and regular review.

Collaboration between industry and policymakers is key when it comes to AI to help overcome the potential for fragmentation and friction, and strengthen understanding, quality, and efficiency. Indeed, regulatory cooperation is increasingly becoming a precondition for effective regulation.

Competing priorities

AI has the power to be transformative for companies, government, and society — but it comes with risks. Policymakers should balance these two competing realities, limiting the potential harm that could arise from unchecked development and implementation of AI without stifling its innovative potential.

Speed of change

Technology moves fast, and faster than governments typically enact related policy. This means a *reactive* approach to mitigate issues created by new technology is unlikely to lead to optimal outcomes.

Added to this, AI's speed and power means that existing issues, inequities, or biases that already exist in societies may be exacerbated by the widespread use of AI. Without appropriate regulation, oversight, and cooperation between government and industry, AI may have the unintended consequence of scaling these inequities. Proactive issue identification and prevention is thus critical to the technology's success.

Better regulation

Policy makers need a proactive and flexible approach to AI regulation that can manage this complex landscape. Deloitte's research suggests five key elements for consideration:

- **Adaptive regulation:** Shift from "regulate and forget" to a responsive, iterative approach.
- **Regulatory sandboxes:** Prototype and test new approaches by creating sandboxes and accelerators.
- **Outcome-based regulation:** Focus on results and performance rather than form.
- **Risk-weighted regulation:** Move from one-size-fits-all regulation to a data-driven, segmented approach.
- **Collaborative regulation:** Align regulation nationally and internationally by engaging a broader set of players across the ecosystem.

Further information can be found below:

- William Eggers, Mike Turley and Pankaj Kamleshkumar Kishnani, "The future of regulation - Principles for regulating emerging technologies", Deloitte, June 2018.
- William Eggers, Mike Turley and Pankaj Kamleshkumar Kishnani, "The regulator's new toolkit - Technologies and tactics for tomorrow's regulator", Deloitte, October 2018.
- Beena Ammanath, "Investors are pouring billions into artificial intelligence. It's time for a commensurate investment in A.I. governance" Deloitte, January 2023.

Review of papers

How the US government can accelerate AI entrepreneurship

August 2022 | Link | Tasha Austin, Deloitte US and Kevin Lubin, Deloitte US.

Summary of key points

The US government has made innovation in AI a top priority, and entrepreneurs and small business have historically been critical to technological innovation. However, these stakeholders face several critical challenges, including:

- **High cost of computing:** Training AI models requires significant resources – specifically, computational resources – and can be cost prohibitive for newer or smaller firms.
- **Low-quality data:** Data cleaning is time- and resource-intensive. Without high-quality data, innovators cannot train their algorithms to reach thresholds necessary to be commercially viable.
- **Scarcity of talent:** Some 250,000 data scientist jobs are estimated to be unfilled. Small businesses must compete with large, established tech companies to access an already limited labor pool.
- **Hurdles to government contracting:** Research indicates that technology startups produce "more disruptive and impactful innovations" when they are able to access public funding, such as federal contracts. However, the US federal contracting system can be confusing and expensive to navigate, thus discouraging smaller companies from engaging.

The authors argue that to ultimately advance AI innovation in the US, greater support is needed for AI entrepreneurs and small businesses.

Policy implications

The authors specifically call on the US government to utilize three key functions – buyer, regulator, and infrastructure provider – to build an ecosystem in which emerging AI firms can thrive. Doing so will likely have positive impacts for the entire technology ecosystem.

Specifically, the authors recommend that the US government:

- **Directly fund AI innovation,** whether through partnerships with universities, spending on pilot projects, or use of contracting dollars.
- **Write, update, and pass legislation that streamlines existing policies and regulations and creates new incentives** for the AI entrepreneurship ecosystem, including by assessing work visas, tax incentives, data flow protections, and contracting requirements.
- **Provide the infrastructure for advanced AI development**, such as through regional hubs, via access to low-cost computing infrastructure, and cross-sector consortiums.

AI for smarter regulation

June 2022 | Link | Adapted from testimony to US Congress by Joe Mariani, Deloitte US.

Summary of key points

The author argues that governments can improve their legislative processes through the application of AI. The article leverages studies on the impact of AI across government – saving workers time and leading to agency AI adoption across the United States federal, state, and local governments – to make two primary recommendations to US lawmakers regarding the use of AI.

- **Assessing the impact of existing legislation:** The scope and volume of legislation is a challenge for humans. The author argues that makes it an "ideal challenge" for AI, noting that machine learning models could find patterns in public policy, which policymakers could then use to appropriately target legislation. The author notes, however, that humans will still need to determine *what* outcomes are considered successes or failures and whether any overall benefit is worth the cost.
- **Assessing the impact of future legislation:** The author notes the importance of simulators in major scientific endeavors, such as US space missions. Training an algorithm on historical data and using it to project trends, for example, could provide a view of trending dynamics in an industry sector. Again, the author cautions that humans may still have to make value judgements – what is the optimal choice given existing values and assumptions?

Policy implications

This paper argues that while human-machine teaming has promise, carefully mitigating risks related to data quality, security, and workforce concerns are critical to an effective outcome.

- **Data and model governance:** AI outputs depend on robust and reliable models and data. For example, the article suggests that organizations should carefully assess the data required for AI operations, consider utilizing open public data to bolster reliability, and tag data with appropriate use cases to ensure it is used only in the appropriate contexts. Similarly, attention should be paid to the transparency of assumptions so that human teams can better understand the context in which AI simulations reach their conclusions.
- **Security:** Beyond typical cybersecurity considerations, AI models utilized for policymaking should be carefully safeguarded.
- **New processes, new skills, new training:** Policymakers should prepare to consume, understand, and question the outcomes that algorithms can provide.

Earning digital trust: Where to invest today and tomorrow

February 2022 | Link| Deborah Golden, Deloitte US, Jesse Goldhammer, Deloitte US, Jay Parekh, Deloitte India, Curt Aubley, Deloitte US, Michael Morris, Deloitte US, Diana Kearns-Manolatos, Deloitte US.

Summary of key points

This paper argues that organizations should invest to earn "digital trust" – protecting their data and information to safeguard relationships, reputation, and revenue. Surveys have found that consumers lose trust in brands after a breach. Investments and strategies to prevent such a breach and loss of trust are thus vital. Organizations should address digital trust through an end-to-end interdisciplinary approach, using technology as an enabler.

This article conducts interviews with more than a dozen specialists to identify four potential technology solutions for advancing digital trust: AI-based data monitoring, cloud-enabled data trusts, blockchain, and quantum technologies.

The authors argue that two of these technologies are ripe for use now:

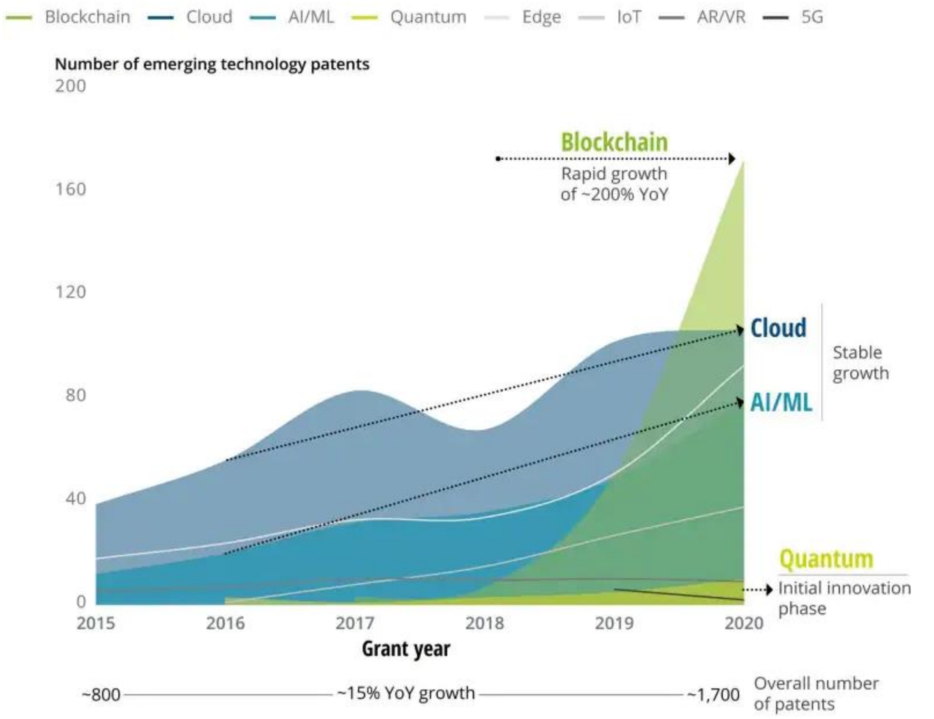
- AI can help to ensure data is correct, complete, and secure, and it can monitor usage to ensure data is used as intended. It can also improve identity and access management by protecting from unauthorized access and detecting irregular behavior. At the same time, AI does present some challenges, including the challenge of active or passive bias.
- Data trusts provide a method for validating, controlling, securing, and sharing information, governing the data's overall use. Data trusts can range from a single entity to a collective, and they largely ease challenges around data management and sharing while also adding privacy and data protection. Cloud technology makes data trusts even more effective.

Two other technologies could transform digital trust in the future. The authors argue that both should be on organizations' radar given their potential.

- Blockchain provides a mechanism to trust the details that it imparts into an independently verifiable and unchangeable database or ledger. Blockchain can help maintain a trusted record of transactions, supporting the ability of users or consumers to access and trust information from a supplier. Blockchain also has benefits for being able to validate trust identities, establish asset ownership, and enable faster legal agreements. Blockchain is not as mature as other technologies, though the authors expect that it will advance in the coming years and result in a potentially transformative impact on digital trust.
- Quantum technologies are expected to have several impacts on digital trust. Immense computing power, when applied to cyber or privacy data, will have the ability to detect suspicious behavior quickly. These technologies may offer enhanced cybersecurity elements. However, they may also provide the ability to render less safe common encryption techniques, making some data and transactions more vulnerable.

Data trust is not just an issue for an organization's CISO or CIO; these authors argue that it also requires engagement, investment, and dedication from the CEO and other business leaders.

Digital trust patents granted annually between 2015 and 2020



Notes: All information on patents is sourced from Derwent World Patents Index via Quid (https://quid.com). The purpose of the analysis is to identify general themes in digital trust. Deloitte did not review any individual patents in preparing this analysis.

Source: Deloitte analysis.

Deloitte Insights | deloitte.com/insights

Policy implications

Many of the technologies touted by the authors are subject to minimal regulation and are still developing. Many – though not all – policymakers are not yet fluent on rapidly changing technologies (i.e., blockchain and quantum computing), leaving their regulation in flux.

The 5 key challenges of AI governance and our learnings

February 2022 | Link | Michelle Lee, Deloitte UK and Andy Whitton, Deloitte UK.

Summary of key points

The authors put forward five key learnings about AI governance:

- Regulations and consumer expectations may vary across countries.
- Careful consideration should be given to what governance should be mandatory and standardized versus what should be discretionary and specific to the use case.
- A clear and well-understood approach to managing third party technologies is essential.
- Training and awareness across the three Lines of Defense (“3LOD framework”) is vital to effective governance.
- Tools and methods for monitoring and testing AI are critical to addressing the challenges presented.

Policy implications

While many organizations using AI have expressed their preference for common rules and policies, this desire will need to be balanced with consumer expectations, which may differ. It is possible myriad policy conversations across geographies results in country- or use case-specific requirements, and it remains to be seen whether globally applicablestandards can be agreed upon to serve as a foundation.

AIs wide shut: AI regulation gets (even more) serious

December 2021 | Link | Duncan Stewart, Deloitte Canada, Paul Lee, Deloitte UK, Ariane Bucaille, Deloitte France, and Gillian Crossan, Deloitte US.

Summary of key points

The authors argue that regulation may finally be catching up with the speed of technological innovation and predicts that 2022 will have large amounts of discussion on regulating AI. The authors surmise that enactment will not take place until 2023 or beyond.

The authors lay out several data points for this trend, including that AI grows vastly more powerful, capable, and affordable than in previous years; regulators have concerns about AI's impact on fairness, bias, discrimination, diversity, and privacy; and AI regulation is a "competitive tool" globally and the first country or region to set standards may have a competitive advantage.

Policy implications

The authors argue that stakeholders ranging from AI users, vendors, and regulators will all be impacted by this ecosystem. They put forward a series of scenarios for the future:

- Certain AI-enabled **features are unavailable** in some geographies, or continue to operate but are subject to fines.
- Major markets implement **conflicting AI regulations** that make it challenging for companies to comply.
- One type of AI regulation emerges as the **"gold standard,"** which could simplify cross-border compliance.
- **Self-regulation** among AI vendors and platforms. In this case, however, the authors argue that regulators are unlikely to completely step aside.

Trustworthy open data for trustworthy AI

December 2021 | Link | Tasha Austin, Deloitte US, Kara Busath, Deloitte US, Allie Diehl, Deloitte US, Pankaj Kamleshkumar Kishnani, Deloitte, India, and Joe Mariani, Deloitte US.

Summary of key points

This paper argues that explainability is a key element to mitigating risk associated with AI deployment. Explainable AI (XAI) refers to techniques that assist the developer in adding transparency to demonstrate how an algorithm produces its output. These tools can help organizations be responsible for questions from consumers and users on how algorithms work and come to the conclusions they do – eliminating the view of AI as a "black box" (e.g., if AI is used to recommend a certain medication to a patient, but does not provide sufficient transparency, the patient may not trust the AI or the recommendation).

Policy implications

The authors put forward the importance of explainability in all AI models, and the role of XAI in building trust. It follows, therefore, that organizations utilizing AI (including government agencies) should prioritize XAI in their systems and policies.

To the degree that governments are regulating AI and algorithm use, these authors would be interested in how to ensure explainability.

The value of XAI is especially noteworthy in what the authors call "sensitive" use cases. This concept is likely to show up in global AI regulatory efforts – i.e., the EU's "high risk" classification proposal for certain AI applications.

Keeping AI Private: Homomorphic encryption and federated learning can underpin more private, secure AI

December 2021 | Link | Duncan Stewart, Deloitte Canada, Ariane Bucaille, Deloitte France, and Gillian Crossan, Deloitte US.

Summary of key points

For major corporations, security in AI is a major concern. Homomorphic Encryption (HE) and Federated Learning (FL) are two emerging technologies that are making AI integration safer, easier, and more focused on securing the privacy of users. Personal information and data leaks are a consistent downfall of cloud-based operations for major corporations; when servers are hacked, cloud data is stolen and personal security information, data and financial information are vulnerable to misuse.

As a result of focusing on security within the AI industry, market growth of HE and FL is predicted to reach US$500 million by 2025. HE allows machine learning to use data while it is encrypted; all other machine learning needs to decrypt the data first, making it vulnerable. FL distributes machine learning to local or edge devices rather than keeping all the data in the same place where one hack could expose it all, which is the case with centralized machine learning. They are not mutually exclusive: HE and FL can be used at the same time.

Companies that continue to use AI in their day-to-day operations are looking at HE and FL as a way to reduce the future risks of stolen data and personal information.

Policy Implications

Within the public sector, organizations are interested in exploring the applications of HE and FL in sensitive industries such as healthcare, finance, civil service, and the justice system. This will likely have positive impacts on the safety of personal information and security of public services and structures. Within the private sector, adopting HE and FL technologies will enable and encourage greater information sharing between organizations without the fear of exposing private intellectual property or exposing personal data collected from clients, customers and partners. In summary, private, secure AI could result in improvements in strategic, operational, and competitive positioning. The benefits for individuals, businesses and society will likely continue to grow with effective security and protection of data.

AI model bias can damage trust more than you may know. But it doesn't have to.

December 2021 | Link | Don Fancher, Deloitte US, Beena Ammanath, Deloitte US, Jonathan Holdowsky, Deloitte US, and Natasha Buckley, Deloitte US.

Summary of key points

Executives report high spending – some 68% surveyed reported that their group invested US$10M or more in the past fiscal year in AI projects – making it a critical investment worth protecting.

At the same time, AI model bias is both prevalent ("more prevalent than many organizations are aware") and highly damaging to trust – of customers, employees, and the general public.

AI model bias usually occurs when the data used to *train* an AI algorithm is not accurate or reflective of reality.

The article identifies two types of bias – passive (not the result of a planned act) and active (caused by human action). Both can occur without intent:

- Examples of passive bias include selection bias (over- or under-including a group, including through insufficient data or poor labelling); circumstantial bias (when changing circumstances or other factors make training data inaccurate); and legacy or associational bias (when AI models are trained with data associated with legal based on certain characteristics, even if unintentionally).
- Examples of active bias include adversarial bias (nefarious "poisoning" of data) and judgement bias (bias is introduced by the misapplication of an AI's decision output).

Bias in AI presents a series of costs: to fix the technology, in lost productivity or revenue, reputational harm, investment loss, and impacts to staffing. A long-term loss of trust can "prevent a company from fulfilling its goals and purpose."

To mitigate the potential costs of model bias, organizations should be proactive in "understanding, anticipating, and… avoiding" bias; doing so will help preserve trust.

The article puts forth four recommendations for organizations, including education within the organization about the potential for AI model bias risk, establishing a "common language" to discuss risk and how to mitigate it; integrating individuals impacted by the model into the model's development; and integrating technology and process into the solution.

According to the authors, AI and trust are "inseparable" – there cannot be trust when relying on flawed AI models. Even a flawless AI does not matter if it exists in an environment lacking trust.

Policy implications

The paper includes examples of AI advancing bias and damaging trust, including in financial services and banking, recruiting and hiring, health care and health insurance, law enforcement and criminal justice, and even college acceptance processes.

It also suggests that bias is "domain agnostic." The article finds that bias exists not only in customer-facing, external models, such as those that make the headlines. Bias

A methodical approach to treating bias

Identify bias in your dataset on a variety of fairness metrics.

Mathematical definitions of fairness cannot all be simultaneously met. Discover in which fashion your model is biased along any dimension.

Investigation of why these biases exist

Quantitative analyses of potential proxies to protected features along with an assessment questionnaire to identify biases along the lifecycle.

Optimize models for both performance and fairness

Explore the trade-offs between the key fairness indicators and key performance indicators.

Compare key performance indicators and key fairness indicators across models

Compare iterations of the same model to track performance vs fairness.

Communicate findings

Automated reporting with key bias risks flagged from the quantitative and qualitative assessments.

is equally prevalent and dangerous in internal, "back office" models that might receive less scrutiny and therefore go undetected for longer periods.

The authors argue that many cases of model bias are not anticipated by organizations working with AI models. The article cites a Deloitte study, ***State of AI***, in which the majority of respondents believed that their models will be fair and impartial (though this article suggests that these respondents may be "oblivious" to the danger of bias).

Investing in trustworthy AI

July 2021 | Link | Kate Schmidt, Deloitte US and Matt Furlow (US - External).

Summary of key points

This paper articulates the challenges and opportunities of AI, beginning with a trust gap among individuals previously surveyed in the United States. Without greater trust, AI's adoption (and its benefits) will likely be slow, and potentially never fully realized. As a result, the authors find that trustworthiness is key to the advancement and deployment of AI.

Survey respondents believed that consumer trust in AI would grow the most when individuals see the personal benefits from adopting AI technologies. They also believed that confidence in AI would grow as it improves individuals' day-to-day work and opportunities (e.g., health and safety, automation of repetitive tasks, and enablement of higher-value and higher-wage growth). The report identifies several commonly cited economic and social benefits of greater AI deployment:

- Improved speed and accuracy of decision-making (e.g., cyber protection, automation)
- Removal or mitigation of bias and subjectivity from high-impact decisions (e.g., hiring and talent management, credit ratings, vendor selections, and higher education admission decisions)
- Faster innovation and pace of discovery (e.g., of medicines, materials, and technologies)
- Increased scale of operations through deployment of fully or partially autonomous agents (e.g., robotics for transportation, production, and delivery)
- Increased ability to detect patterns or anomalies in complex data sets (e.g., fraud detection, health care tracking, utility or service reliability)
- New types of work and specialized occupations focused on creating, managing, and maintaining AI systems (e.g., designing and training algorithms, validation, work for which AI systems augment or speed delivery)
- Reduction of repetitive tasks (e.g., chatbots for customer service, automation of routine formatting of documents, planning and scheduling)
- Increased wages or improved working conditions, including by allowing workers to migrate to higher-value tasks by improving productivity
- Improved safety (e.g., monitoring hazardous conditions, equipment, environmental conditions)

At the same time, survey respondents also believed that bias, lack of human accountability, and a lack of algorithmic explainability all contribute to decreased trust in AI. Additionally, there are concerns from respondents about potential job loss due to increased AI-enabled automation as well as the acceleration of social or economic divides between workers with and without AI skills.

Left unaddressed, the report argues that these challenges in trustworthiness will inhibit the long-term growth and adoption of AI technologies, and the economic and societal benefits that this transition is expected to bring.

Collectively, the report identifies a series of similar ideas around fairness, transparency, and accountability in the development and use of AI applications as “Trustworthy AI,” which it argues is critical to the US national strategy for AI and for realizing all its benefits. Deloitte’s six dimensions of Trustworthy AI (see page 6 for more detail) are:

- Fair and impartial
- Transparent and explainable
- Responsible and accountable
- Robust and reliable
- Respectful of privacy
- Safe and secure

The paper also puts forward solutions to help enable AI trustworthiness, concluding that trustworthy AI can enable innovation and support US global leadership.

The report concludes with two case studies: the first on research and development (R&D) investments, and the second on government modelling of trustworthy AI.

- Case study one models the impact of increased government R&D in AI, citing potential returns on investment of between 6.8% and 9.6%. It also makes the case for second- and third-order impacts on the US economy that could have an impact of as great as US$1.4 trillion of additional GDP through 2025.
- Case study two maps opportunities for the federal government’s promotion of trustworthy AI across four categories: AI implementation in e-government applications that serve citizens, applications used by agencies internally, development of novel AI applications for public crises, and the establishment of procurement processes for AI that establish guidelines for trustworthiness.

Policy implications

The paper makes a case for the United States to lead global technological advancement on AI, but also to lead in *values* – using US influence to ensure trustworthiness is a core component of AI development and deployment globally.

The paper argues that advancing trustworthy AI will enable innovation and support US global leadership and competitiveness. This is a critical part of the US policy conversation – particularly US competition vis-à-vis China and the role of emerging technologies like AI and others in this competition – and it is linked to discussions related to public funding for research and development, high-tech manufacturing and production incentives, and research security, among others.

Specifically, the paper cites the findings of a Congressionally-established committee, the US National Security Commission on Artificial Intelligence (NSCAI), and its finding in March 2021 that a lack of confidence in AI systems would jeopardize long-term US technological competitiveness. It goes on to cite the report's argument:

> "If AI systems routinely do not work as designed or are unpredictable in ways that can have significant negative consequences, then leaders will not adopt them, operators will not use them, Congress will not fund them, and the American people will not support them."

The paper also identifies recommendations from its survey of AI leaders, proposing a handful of policy solutions – namely, increased US federal government investment in AI – that the authors believe would enable innovation and economic growth while mitigating risks posed by AI. The report argues that the US government not only has a critical role to play in promoting the deployment of AI, but that it also has a vital role in managing real and perceived *risks* (and thereby boosting AI trustworthiness) that will ultimately result in greater adoption, use, and benefit to US society and societies worldwide.

The paper's survey finds that respondents strongly believe that public policy – specifically in the United States – can support AI innovation and have positive knock-on effects for economic growth, health, safety, and well-being in the United States. This includes supporting AI-related benefits (e.g., using AI to increase productivity, make more accurate and faster decisions, remove bias from certain processes), as well as mitigation of AI-related risks like bias, accountability, and transparency. Survey respondents also believed that government can contribute to increased worker trust in AI technologies, through processes like worker safety, hiring and talent development, and increased worker efficiency and reduced repetition. More than half of respondents felt that government could support AI adoption to expand access to higher-value work, leading to higher wages and/or better working conditions for individuals.

More than half of survey respondents agreed with the following types of government interventions:

- Increased government investment in AI research and innovation, particularly the earliest stages of AI development
- Increased access to government data sets for training and process improvement

Extract of survey results

63% of respondents reported low outcomes in their organization's AI quality and risk management process and frameworks to assess AI model bias and other risks.

94% of business leaders agreed that AI is critical to success over the next few years.

79% of workers were not educated properly by their organizations on how to use AI effectively.

25% of organizations surveyed provided user-friendly AI systems in the workforce.

50% identified AI-related risks as a significant challenge to scaling their AI initiatives.

79% of leaders reported full-scale deployment for three or more types of AI applications within their organizations.

46% indicated difficulties in integrating AI into their organization's daily operations and workflows.

82% indicated their employees believe that working with AI technologies will enhance their performance.

29% indicated lack of technical skills as a top challenge in starting and scaling AI projects.

- Retraining or continuing education programs for adults
- Encouraging industry-led, consensus-based standards for algorithmic performance and reliability
- Supporting or establishing international partnership to promote common frameworks for AI use and development
- Supporting student curricula that will promote AI skills and careers

In addition to the policy implications of this paper's survey, the authors put forward a series of specific policy recommendations. These include:

- **Standards:** Supporting the development of AI trustworthiness standards.
- **Resources:** Leveraging federal resources to accelerate innovation, including research and data-sharing. The report's authors specifically identify access to government data sets for private sector research as a priority, as well as enabling shared computing resources. The report also identifies open-source tools and frameworks as a method for encouraging government, academia, and private sector cooperation.
- **Partnerships:** Creating or supporting international partnerships that promote trustworthy AI development and deployment, including through US global standards. This recommendation may have implications for future US digital trade agreements and participation in global standards-setting fora in which global "rules of the road" for technologies like AI are debated and developed. The Biden administration has generally supported US-led technology standards (especially visible, for example, on 5G).
- **Government use:** Modeling responsible AI implementation through government applications, including by federal agencies, and the application of AI to timely and relevant events, such as COVID-19 and climate change. This recommendation also identifies the establishment of procurement processes focused on trustworthy AI as a potential policy solution.
- **Standards:** Supporting US workforce development of AI-related skills, including for adults who need to reskill to transition to an AI-related job, and for students who could benefit from the promotion of AI skills and careers.

While focused overwhelmingly on US AI development, deployment, and policy, this report notes implications for other geographies:

- **China:** China has invested heavily in its domestic AI industry. The report calls for the United States to counter China's "digital protectionism" and its active intervention in commercial development of AI.
- **Russia:** Russia has been more "explicit" in viewing AI as a competitive advantage for its military technologies – an especially visible goal in the United States, as policymakers respond to Russia's invasion of Ukraine by enacting stringent sanctions and export controls focused on Russia's ability to develop, manufacture, and procure high-tech components that could be used for military purposes.
- **European Union:** Draft EU regulations consider certain AI applications "high risk" and potentially subject to certain restrictions. As with European privacy regulations, this proposal would be extraterritorial and thus influence the development and deployment of US AI technologies. The report calls on US policymakers to balance privacy and data protection concerns with negative impacts to competitiveness in its response to the European Union and other regulation.

How to spot unintended bias in machine learning

March 2021 | Link | Michelle Lee, Deloitte UK.

Summary of key points

Unintended bias in machine learning (ML) can lead to discrimination and exacerbate existing societal inequities. While a series of frameworks for mitigating bias exist, they have not been widely operationalized into practical tools, this paper argues.

Six types of bias

Historical bias – a misalignment between the world "as it is" and the values or objectives required from the model (occurs in selection, measurement, and pre-processing stage of ML).

Representation bias – the under-representation or failure to generalize of a group in the population (occurs in population selection stage of ML).

Measurement bias – choosing poor proxies for real-world quantities (occurs in data measurement selection stage of ML).

Aggregation bias – improper combination of distinct groups into a single model (occurs in model training stage of ML).

Evaluation bias – improper performance metrics or testing/benchmarks that are not representative (occurs in model evaluation stage of ML).

Deployment bias – improper use or interpretation of a model (occurs in outcomes processing stage of ML).

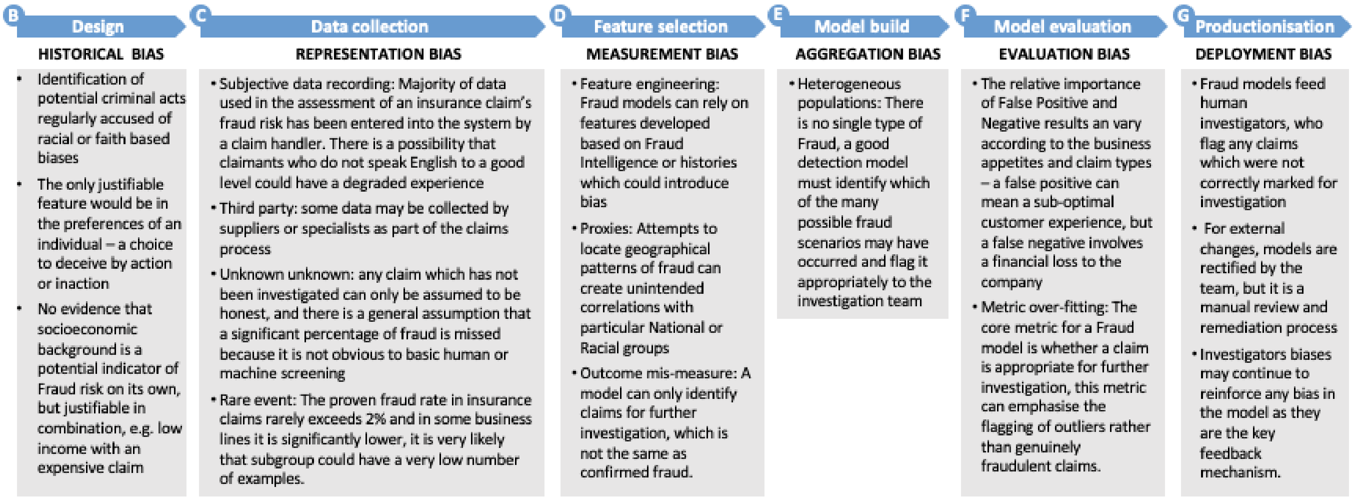
Managing risk is possible through best practices, though there has not yet been a systematic effort to integrate them into organizational risk management processes (RMP).

Additionally, most frameworks lack practical, operational steps for integrating these best practices. This paper introduces a bias risk identification questionnaire to help detect bias in each stage of ML development. The questionnaire is intended to identify both *how* a model might be biased, as well as *why*, to better manage said bias.

Policy implications

The paper includes a case study that applies the article’s questionnaire to a case of predicting insurance fraud. Taking this case study through the questionnaire suggests opportunities for mitigation.

use



B Design	C Data collection	D Feature selection	E Model build	F Model evaluation	G Productionisation
HISTORICAL BIAS	REPRESENTATION BIAS	MEASUREMENT BIAS	AGGREGATION BIAS	EVALUATION BIAS	DEPLOYMENT BIAS
• Identification of potential criminal acts regularly accused of racial or faith based biases • The only justifiable feature would be in the preferences of an individual – a choice to deceive by action or inaction • No evidence that socioeconomic background is a potential indicator of Fraud risk on its own, but justifiable in combination, e.g. low income with an expensive claim	• Subjective data recording: Majority of data used in the assessment of an insurance claim’s fraud risk has been entered into the system by a claim handler. There is a possibility that claimants who do not speak English to a good level could have a degraded experience • Third party: some data may be collected by suppliers or specialists as part of the claims process • Unknown unknown: any claim which has not been investigated can only be assumed to be honest, and there is a general assumption that a significant percentage of fraud is missed because it is not obvious to basic human or machine screening • Rare event: The proven fraud rate in insurance claims rarely exceeds 2% and in some business lines it is significantly lower, it is very likely that subgroup could have a very low number of examples.	• Feature engineering: Fraud models can rely on features developed based on Fraud Intelligence or histories which could introduce bias • Proxies: Attempts to locate geographical patterns of fraud can create unintended correlations with particular National or Racial groups • Outcome mis-measure: A model can only identify claims for further investigation, which is not the same as confirmed fraud.	• Heterogeneous populations: There is no single type of Fraud, a good detection model must identify which of the many possible fraud scenarios may have occurred and flag it appropriately to the investigation team	• The relative importance of False Positive and Negative results an vary according to the business appetites and claim types – a false positive can mean a sub-optimal customer experience, but a false negative involves a financial loss to the company • Metric over-fitting: The core metric for a Fraud model is whether a claim is appropriate for further investigation, this metric can emphasise the flagging of outliers rather than genuinely fraudulent claims.	• Fraud models feed human investigators, who flag any claims which were not correctly marked for investigation • For external changes, models are rectified by the team, but it is a manual review and remediation process • Investigators biases may continue to reinforce any bias in the model as they are the key feedback mechanism.

Adapted from: Lee, Michelle Seng Ah and Jatinder Singh. “Risk identification questionnaire for unintended bias in machine learning development lifecycle.”

The most significant implications of this article are not public policy focused, instead it highlights corporate governance and internal processes that can be used to prevent, identify, and mitigate bias. However, the questions posed at a company level are also relevant at a policy level, primarily for identifying and understanding the bias risk from the use of AI, then mitigating such risk as much as possible. For the most part, policymakers have not yet advanced regulations that achieve these goals, though it remains an area of active discussion across the United States and Europe.

Striving for fairness in AI models

March 2021 | Link | David Thogmartin , Deloitte Denmark, Andy Whitton, Deloitte UK, and Michelle Lee, Deloitte UK.

Summary of key points

This paper argues that no single algorithm can account for all the elements of "fairness," which itself is a complex definition without consensus. Given mathematical definitions of bias can contradict one another, bias must be evaluated in ways that are compatible with the particular use case for AI in order to achieve "fairness."

Bias is extremely complex and "practically unavoidable," meaning that decision models should strive to balance fairness with optimization of the model. Removing "protected features" from a model (race, gender, religion, age, etc.) will likely not remove bias, as models utilize proxy features for clues that lead back to protected class.

As a result, the authors argue that optimizing for fairness and performance is an ongoing task – populations and data change, as can the model's efficacy.

Policy implications

This paper lays the groundwork for understanding why governments around the world are focused on defining important issues like bias and fairness and integrating them into AI regulations. But it also suggests that any regulation – to be truly effective – must be flexible and dynamic enough to accommodate a changing and complex issue like fairness.

At the same time, the paper presents human-caused errors that can result in biased algorithms. As much as possible, this paper indicates that systems should be put into place to root out those errors – a key focus of legislators and regulators thinking about issues in AI.

Why technology cannot solve algorithmic fairness: gaps between how computer scientists and ethical philosophers define fairness

November 2020 | Link (Part 1), Link (Part 2), Link (Part 3) | Michelle Lee, Deloitte UK.

Summary of key points – Part 1

The increased use of machine learning to inform critical decisions has led to a concern about issues of bias. Scholars have created mathematical tools to test these algorithms on a pass/fail basis against a number of mathematical definitions of fairness.

However, according to the author, it is mathematically impossible to meet all these fairness conditions simultaneously. Therefore "fairness" cannot be reduced to a formula; it depends on context. In fact, automating "fairness" is incompatible with EU non-discrimination law.

There is a gap between mathematical definitions of "fairness" and philosophical/welfare economics definitions. Formulas based on the egalitarian foundation that everyone should be treated equally, do not align with the politics/philosophy of some inequalities being acceptable to society.

Summary of key points – Part 2

This paper dives deeper into “fairness” in the context of what is within an individual’s control. Philosophies recognize a difference between “effort” and “circumstance,” but in reality, splitting these out in a ML model can be difficult. For example, race and gender may be causally relevant in differential medical diagnosis (e.g., sickle cell anemia, ovarian cancer).

Feedback loops exist that need to be considered. For example, the paper uses as an example a history of poor credit that may lead the algorithm to offer less credit. This may be considered unfair, but trying to mitigate the bias may make matters worse. Giving credit to those who can’t afford it and have a higher chance of default, will only make their credit rating worse.

According to the authors, using welfare economics to justify inequalities – the idea that individuals bear the consequences of their choices – is not “fair” if individuals do not have access to the same choices.

Focusing on narrow bias factors misses the impact on an individual’s welfare and autonomy. The narrow definition of unfair bias in each of these metrics only provides a partial snapshot of what inequalities and biases are affecting the model and does not consider the long-term and big-picture ethical goals beyond this equalization.

Summary of key points – Part 3

Similar to how a company may define a set of quantifiable values to gauge its achievements using Key Performance Indicators, the authors argue that there should be outcome based, quantifiable statements from an ethical standpoint: Key Ethics Indicators, enabling developers to manage and track to what extent each model is meeting the stated objectives.

Additionally, this paper suggests that the roles and responsibilities of a developer are necessarily intertwined with the role of the expert or business stakeholder, as the ethical and practical valuations of what “success” looks like in the model directly influences the algorithm design, build, and testing.

Policy implications

This paper argues that the concept of “fairness” is subjective and context-dependent. Bias should be addressed and minimized, but the optimum way to do this is dependent on political/philosophical choices.

Therefore, it is possible that blanket policies to address bias will be blunt instruments that may have unintended consequences. Instead, this paper suggests that the issues of bias should be considered on a case-by-case basis and include a wide variety of stakeholder input.

Risk-based approach to AI ethics: operationalizing values and principles

May 2020 | Link | Michelle Lee, Deloitte UK, Kate Lavinenko, Deloitte UK.

Summary of key points

Companies have started publicizing ethical values, often specific to AI. This approach focuses on how an AI solution should behave to be "good," to enable those who are affected by it to trust in it. Frequently cited values include fairness, accountability, and explainability. A key challenge, according to the authors, is operationalizing these principles into the AI development lifecycle.

Organizations have additionally introduced "ethics by design" principles to embed these ethical values into AI products, similarly to "privacy by design." The challenge with this is different interpretations of what is considered to be ethical/fair.

This paper argues that an effective AI ethics governance requires a risk-based approach. The primary focus is the risk to fundamental human rights and freedoms that are widely accepted and enshrined in the EU Charter of Fundamental Rights and Freedoms, including the right to privacy (Article 7, 8, 10, autonomy, sanctity of home, personal autonomy, communication secrecy), the right to equality (Article 21, equal treatment, prohibition of discrimination), and the freedom of expression (Article 11).

There has been a proliferation of public and private initiatives that describe high-level principles and values to guide ethical AI. Over 65 frameworks have been collected in the AI Ethics Guidelines Global Inventory, collected by external inventory AlgorithmWatch.

Policy implications

This paper's findings suggests that a risk-based approach to governance of AI is needed. The EU is looking to introduce this across Europe where uses of AI that are considered "high risk" are more tightly controlled or even prohibited. However, this paper shows the need to complete risk assessments on a case-by-case basis to identify where ethical values and business objective may be at odds with one another.

"Trustworthy" AI is a framework to help manage unique risk

March 2020 | Link (External) | Irfan Saif, Deloitte US and Beena Ammanath, Deloitte US.

Summary of key points

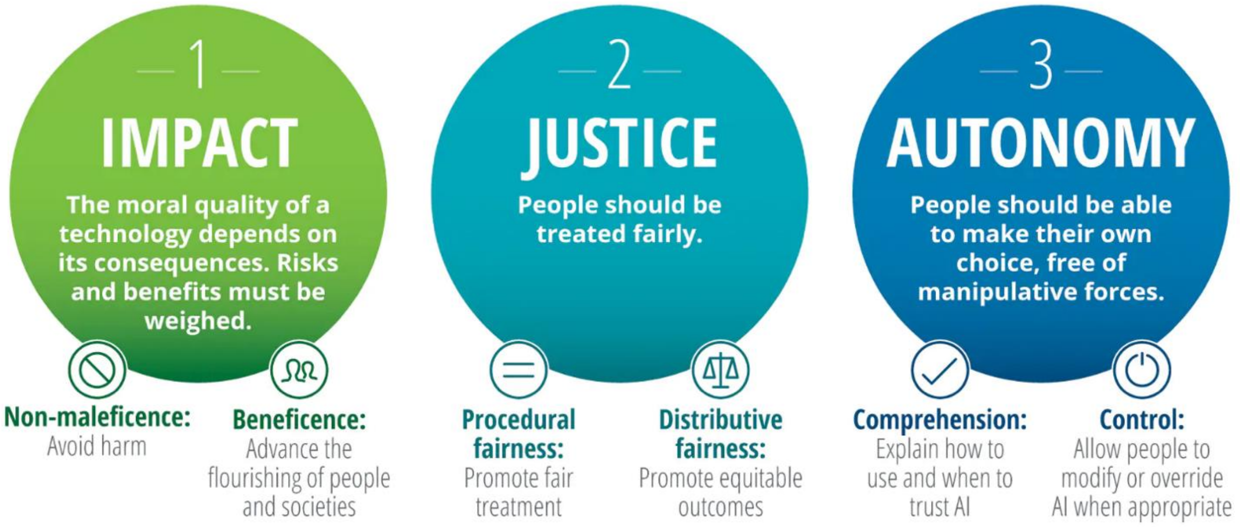
The barrier to widespread AI deployment is no longer the technology, it is ethics, governance, and human values. This paper argues that the risks associated with deployed AI increase as the technology is more widely adopted – ranging from societal impacts (bias, discrimination) to business ones (lawsuits, regulatory fines, angry customers, loss of revenue and reputational damage). AI is also no longer a "nice to have," meaning there is no option to opt out of AI's promise (and associated risks). Instead, the authors argue that organizations need an organized framework to ensure integrity and trust with both internal and external stakeholders.

Deloitte proposes a Trustworthy AI framework, which is designed to help organizations identify and mitigate potential risks related to AI (see page 6 for more detail):

- Fair, not biased
- Transparent and explainable
- Responsible and accountable
- Robust and reliable
- Respectful of privacy
- Safe and secure

Human values in the loop: Design principles for ethical AI

January 2020 | Link | Jim Guszcza, Deloitte US, Michelle Lee, Deloitte UK, Beena Ammanath, Deloitte US.



1 IMPACT

The moral quality of a technology depends on its consequences. Risks and benefits must be weighed.

Non-maleficence: Avoid harm

Beneficence: Advance the flourishing of people and societies

2 JUSTICE

People should be treated fairly.

Procedural fairness: Promote fair treatment

Distributive fairness: Promote equitable outcomes

3 AUTONOMY

People should be able to make their own choice, free of manipulative forces.

Comprehension: Explain how to use and when to trust AI

Control: Allow people to modify or override AI when appropriate

Source: Deloitte analysis.

Deloitte Insights | deloitte.com/insights

Summary of key points

This essay attempts to illustrate that ethical principles can serve as design principles for organizations seeking to deploy innovative AI technologies that are economically profitable as well as beneficial, fair, and autonomy-preserving for people and societies. Specifically, it proposes “impact,” “justice,” and “autonomy” as three core principles that can usefully guide discussions around AI’s ethical implications.

Impact – the moral quality of a technology depends on its consequences. Risks and benefits should be weighed.

- Non-maleficence – safety, reliability, robustness, data provenance, privacy, cybersecurity, misuse. For example, this includes refraining from causing intentional harm through phishing, cyber breaches, weaponized AI or fake news. It also extends to avoiding unintentional harm due to false positives, faulty data, poor model specification, or poor algorithm operationalization.

- Beneficence – human flourishing, well-being, dignity, common good, and sustainability. This includes, for example, using AI to improve medical care, deliver public benefits, create safer environments, or improve educational outcomes.

Justice – individuals should be treated fairly.

- Procedural fairness – algorithmic bias, equitable treatment, and consistency. An example of this includes facial recognition software that recognizes the faces of black people just as reliably as the faces of white people, or internet searches that avoid amplifying implicit social biases.
- Distributive fairness – shared benefits, shared prosperity, and fair decision outcomes. For example, addressing growing inequality due to technology-induced workplace changes, and avoiding algorithmic biases that lead to unfairness in hiring or parole decisions.

Autonomy – people should be able to make their own choices free of manipulative forces.

- Comprehension – intelligibility, transparency, trustworthiness, and accountability. For example, Explainable AI algorithms helping judges or hiring managers make better decisions, a vehicle operator understanding when to trust autopilot technology, an AI-based tool informing decision makers when they are being “nudged,” as well as a chatbot not masquerading as a real human.
- Control – consent, choice, enhancing human agency and self-determination, and reversibility of machine autonomy. For example, ensuring decision-makers (such as a vehicle operator) can override an algorithm that is clearly going astray. Choice architecture enables access to the full menu of choices if the algorithmically generated default isn’t acceptable.

Policy implications

Ethics is often viewed as a constraint on organizations’ abilities to maximize shareholder returns. But this paper suggests a different perspective: ethical principles can serve as design criteria for developing innovative uses of AI that can improve well-being, reduce inequities, and help individuals better achieve their goals.

To do so, this paper suggests that AI models need to consider ethics in their design with appropriate monitoring to make sure any unintended consequences are quickly identified and addressed. However, this is not as simple as a mathematical formula, as discussed in previous papers. Instead, this framework suggests that companies should be open and transparent about the ethical choices they have made and their impact.

Contacts

Beena Ammanath
Executive Director
Global Deloitte AI Institute
Deloitte Consulting LLP

Simon Cleveland
Partner, Public Policy
Deloitte Global

Tim Smith
Principal, Tech Strategy & Business
Transformation Leader
Deloitte US

Diana Kearns-Manolatos
Senior Manager
Center for Integrated Research
Deloitte Services LP

Acknowledgements

Special thanks to the following individuals who made this report possible.

Richard Bailey   Kelsey Lilley   Mosche Orth   Alexia Tassios