ARTIFICIAL INTELLIGENCE & ETHICS

The fundamentals every organizational leader should consider when embracing AI.

A PRIMER FROM SAS

CONTENTS

ETHICS AND AI: WHAT IS ALL THE FUSS ABOUT?

Today, AI is everywhere, touching every waking moment of our lives from the mundane to the quirky, the lifesaving to the momentous. AI is in play as the personal assistants that make our shopping lists, as apps that show us how we'll look when we're old and so much more besides. At the other end of the scale, it's making huge improvements to human lives - such as helping to uncover cancer risks and predicting the impacts of cyclones with 99% accuracy.

AI is even reinventing invention. Consider drug discovery – a slow, expensive process even with the most experienced, knowledgeable chemists. Today, using machine learning, a scientist can explore the 1060 potentially drug-like molecules (more than the atoms in the solar system) that are available[1] in a fraction of the time and at a scale previously unachievable.

In this climate of accelerating AI adoption, ethics have become a critical, emergent concern amongst consumers, business leaders and governments. But why? Simple algorithms–mathematical instructions–date back to antiquity, meaning that humans have been using algorithms to solve problems for decades. So why are AI algorithms raising new challenges around ethical concerns?

Well, AI expands the scope, scale and speed of inferences and actions beyond anything a single human decision-maker could achieve.

The vast data sets, the compute power and the ability to learn and act at scale are characteristics that make AI not just a game-changer, but a new epoch creator because it greatly accelerates the pace and impact of both innovation and the possibility of negative side-effects.

So, what should leaders be aware of when it comes to developing AI and driving an understanding of ethical considerations across the organization?

This paper is designed to provide an overview of the ethics landscape for AI. We will define ethics in the realm of software and illustrate the impacts that factors such as bias can have on outcomes. While acknowledging that there are no watertight ethical decisions, we will touch on how all organizations, including regulators are approaching this important topic.

[1] Technology Review

RISK & REWARD IN AI

The more we rely on AI to inform decisions and actions —either in collaboration with us or autonomously—the greater the concern about whether it will deliver outcomes that align to our intent and to relevant cultural and social mores. As the number of AI use cases continues to expand, we feel its increasing impact on our daily lives. AI can now have a direct bearing on our wellbeing, for example through healthcare provision; our access to resources, such as cheap finance; and even our life opportunities, through recruitment.

It's precisely because of the degree of impact or transformation that AI can have on industries and markets, as well as humans as employees, consumers and citizens, that concern about the scope and effect of AI-driven-decisions is coming to the fore.

This is brilliantly illustrated by the classic example of the autonomous car. It's an ethically sound use of AI—a transport system that can reduce accidents and fatalities while decreasing traffic congestion, fuel consumption and CO_2 emissions.

However, other ethics issues are at play. For instance, while on the road, the vehicle is faced with an urgent dilemma: a pedestrian dashes into the road immediately in front of it.

Does the AI avoid the pedestrian but risk possibly injuring or killing the passengers, or should the AI collide with the pedestrian in order to save the group?

In this scenario there is no outcome that will be ideal for everyone.

In the field of mental health similar subtleties occur. Great work is being done to identify patterns of potential suicide risk.

Using AI for this endeavor is beneficial to humans yet it also generates risk because, by spotting likely victims of suicide, aren't we setting the expectation for intervention? Is it ethical to ask healthcare workers to make those decisions to intervene? If so, do they have the resources to respond effectively and appropriately? What about the rights of the individuals being targeted? Have they agreed to or would they reasonably expect such interventions and from whom?

Even if the solution is deemed ethical, AI solutions are still fallible. Therefore, you will also need to understand your organization's appetite for risk and model the risk versus reward equation for every application of AI. Specifically, it is important to distinguish real risks from those based on misinformation, poor past experiences, or emotional reactions such as fear. The answers to these questions will help determine not just if, but how, your AI solution should be implemented.

KEY ETHICAL CONSIDERATIONS

THERE ARE SEVERAL AREAS TO CONSIDER WHEN EVALUATING WHETHER AN AI SOLUTION IS ETHICAL—IN THEORY OR PRACTICE.

1 DOES THE SOLUTION DELIVER FAIR AND EQUITABLE OUTCOMES?

The objective here is to avoid building systems that create or reinforce inequalities, such as uneven access to healthcare or jobs. Unfortunately, just as human beings can display prejudices—whether conscious or unconscious—so can AI systems that are programmed by humans with certain beliefs or misperceptions. If the data that was used to train the software is partial, or unrepresentative in some way, the problem could be expounded thereby reinforcing current inequalities.

As an example, Amazon was developing an AI system that could help it expedite its recruitment process. The machines were trained to analyze soft copies of the resumes, rating them from one to five, and passing the top five candidates to recruiting team leaders. Unfortunately, the training data was based on 10 years of resumes mainly from men. Therefore, the software taught itself that men were the preference and filtered out women by scanning for female-oriented language in the text—pronouns, use of 'women'—and penalizing those from women-only colleges etc.[2]

In this example, the reinforcement of gender inequality is obvious and explicit, yet there are other situations where the objectives of the AI solution need to be made clearer. Neonatal Intensive Care Units (NICU) provide a great example of a more nuanced scenario. Here, patients are fragile, at high-risk of developing complications resulting in death. Their needs are complex and overall outcomes poor. If, in the interest of determining how to allocate scarce resources, an algorithm is set up solely to optimize cost (i.e. minimize the cost of care relative to the likely outcome), patient outcomes in the NICU would likely plummet.

How and where a solution is deployed can also impact outcomes. Consider an application to evaluate road quality based on the level of jostling recorded by cell phones. Strangely, the poorest road conditions were reported in very expensive neighborhoods. Not because the roads were, indeed, the worst around. Rather, because these were the areas where cell phone usage was predominant at the time. A simple anecdote, but one that illustrates how, when and where a solution is deployed can impact whether the application and resulting outcomes are fair or equitable.

[2] Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, October 2018

2 DOES THE SOLUTION INTRODUCE OR EXACERBATE BIAS?

Bias is often an unfortunate fact of life and is undesirable when it increases inequality or unfairly favors one group over another. Yet bias may be acceptable—in rare circumstances—if mindfully applied to rectify a larger environmental bias already in place, such as socio-economic inequality.

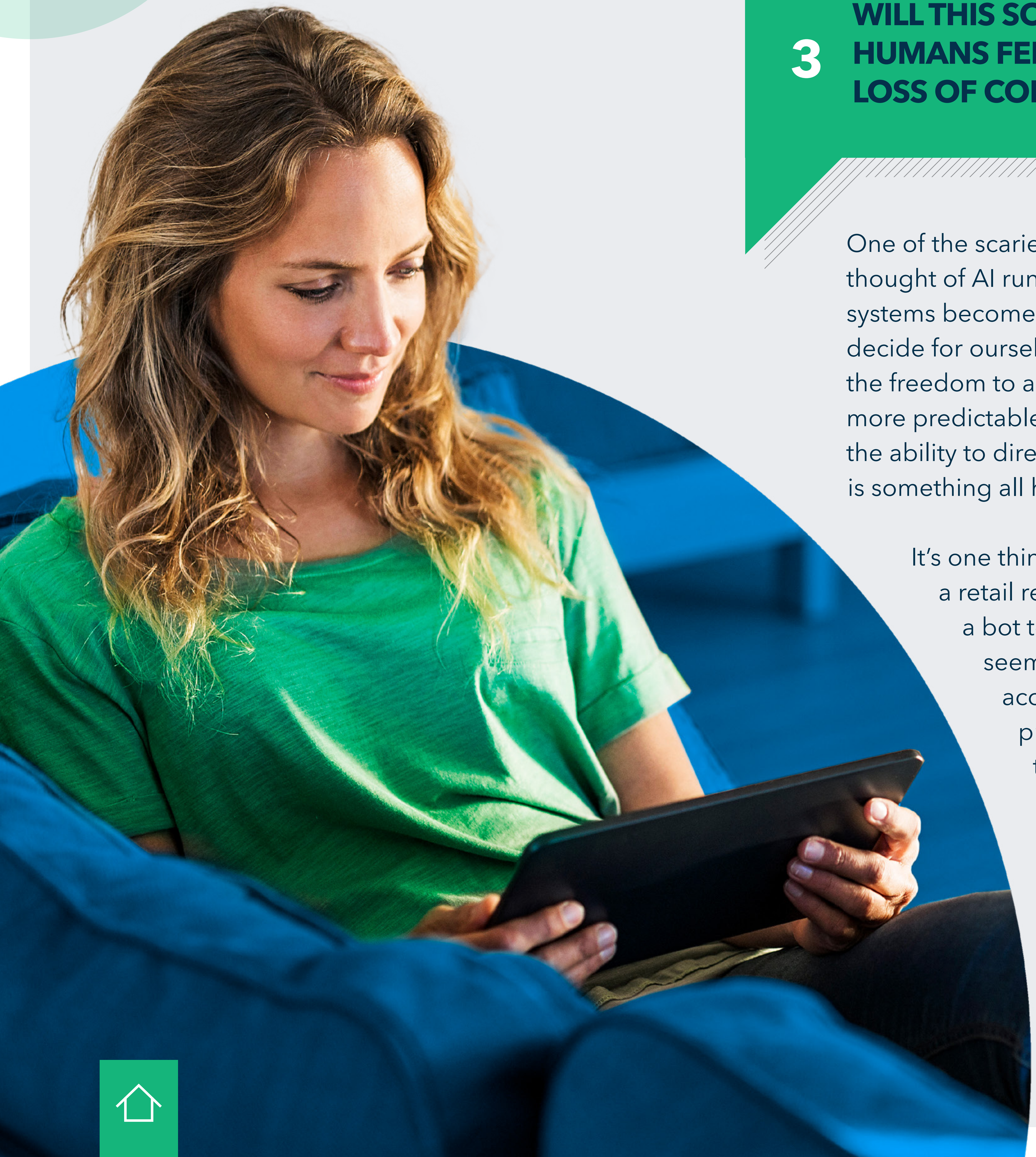
Unfortunately, it is notoriously easy for bias to infect AI solutions. This can occur at three levels—at the data level, in the way data is collected, sampled or selected for use; at the algorithm development level; and/or the deployment level. For instance, even if training data is representative, the humans architecting the model can be unaware of their own prejudices, resulting in bias being missed and their own unconscious biases (of which there are many kinds) impacting AI outcomes.

Left unnoticed, bias can become systematically amplified or reinforced. Let's take the recruitment example again. If an organization hires fewer women, or those they do hire tend to start at lower salaries than similarly qualified male colleagues, this pay trend will likely continue throughout the women's careers. If they are also not promoted as frequently as their male colleagues, who are also a majority in the population, an algorithm using factors such as salary and promotion trends to predict future performance is going to perpetuate that inequity. Even if it doesn't explicitly single out women, women will be disproportionally under-represented in the 'high performing' set based on systemic, historic gender inequalities. This is a good example of systemic bias being perpetuated in direct contrast to the presumably positive intent of the solution in theory.

Similarly, healthcare solutions can lead to ineffective treatment plans for patients' conditions if they are underrepresented in the data used to train the solution —either because those populations are a minority in the actual patient population or they have been historically undertreated or misdiagnosed. Such outcomes can occur even in the absence of nefarious intent.

AI's ability to scale up embedded prejudice is also worrying. Take the example of COMPAS, a recidivism algorithm applied to prisoner populations. It has been used in the US to calculate the likelihood of reoffending—using a score to predict 'likelihood' and as the basis for a judge to make recommendations about judicial outcomes. In Broward County, Florida, this system incorrectly labeled African American defendants as high risk at nearly twice the rate it mislabeled white defendants.[3] A further complexity comes if the algorithm correlated high levels of recidivism to low income. We would still have no evidence that poverty causes crime, yet many of these risk assessments do, in fact, turn correlation insights into causative scoring systems.

[3] How we analyzed the COMPAS recidivism algorithm, ProPublica, 2016



3 WILL THIS SOLUTION RESULT IN HUMANS FEELING OR EXPERIENCING LOSS OF CONTROL OR AGENCY?

One of the scariest consequences for many humans is the thought of AI running our lives. The fear is that as AI-driven systems become more pervasive, we will lose our ability to decide for ourselves, to evaluate alternative options and have the freedom to act. We may all be creatures of habit and far more predictable than we'd like to admit. However, having the ability to direct our own actions without undue influence is something all humans expect to be able to do.

It's one thing for consumers to accept a nudge from a retail recommendation engine or an assist from a bot to draft an email response. Even in these seemingly innocuous cases, we have a choice—accept the recommendation, edit or reject the proposed response. But what about systems that influence or make more fundamental decisions? Or when our opinions and companions are largely determined by algorithmic connection.

As alluded to earlier, will we lose the opportunity to apply individual value-based discretion when there are potentially competing moral dilemmas in play?

To mitigate these concerns, organizations must consider how humans will interact with each AI solution and define the engagement accordingly. Would an individual engaging with the solution reasonably expect it to behave this way? Will the proposed implementation cause individuals to feel or be 'at the mercy' of the algorithm? While consumers have proven to be incredibly open to AI solutions, their acceptance is largely founded on an awareness and agreement that the benefits of the proposed solution outweigh other concerns. With that in mind, clearly communicating when AI systems are in place, how they are used and, if appropriate, allowing people to opt-out, intercede, customize or challenge algorithmic actions or decisions is crucial.

4 WHAT IS THE IMPACT ON EXISTING ROLES AND EMPLOYEES?

We see discussions everywhere about the long-term implications of AI on the future of work. Media speculation has often fanned the flames, driving misperceptions and fear. Even so, it cannot be denied that AI solutions typically offload tasks and change the nature of existing roles. Affected tasks may be onerous or repetitive, but even then, it is human nature to resist change. So, big or small, the impact of AI on existing roles and the need to modify established business practices must be addressed head on. In some cases, existing resources will need to be upskilled or redeployed, though some other jobs may be eliminated.

Beyond changing the nature of existing work, AI will also require employees to become more technically literate.

Consider a diagnostic AI solution that identifies a potential discrepancy in a diagnosis for a specific patient.

While the AI's predictive accuracy may be higher than a human doctor, it is not perfect. Without understanding the context in which the AI was intended to be used and its inherent limitations, a physician cannot properly consider the provided information. Likewise, a driver should be forewarned not to utilize autopilot in heavy rain if the system hasn't been trained to handle this weather condition. While not all (or even most) AI solutions reach this level of life-or-death consequence, each must be evaluated to identify what is required to enable the solution to be appropriately and/or safely utilized.

THE ETHICS MANDATE

Ethics and related risk management topics are no longer just a matter of conceptual debate. As consumers become increasingly conscious and demanding, and regulators step up to the plate, engaging in these discussions becomes a must-have not a nice-to-have. Factor in the increasingly public, incessant and widespread conversations of the future of work, social equity and so on and, companies adopting AI without addressing these concerns do so at their own risk.

THE RISE OF THE CONSCIENTIOUS CONSUMER

Trust is essential in the value exchange that occurs between a consumer/citizen and an organization. Consumers who do not trust how their data is being used and worry that it is being used to disadvantage them in some way, are at risk of withdrawing from the relationship with a brand.

Being found to be deploying AI systems deemed to be unethical can harm your brand's reputation and significantly impact the bottom line. Apple's credit scoring algorithms, for its credit card offered by Goldman Sachs, were found to offer higher credit limits to working spouses than their non-working spouses, who were predominantly female. This would not have generated beneficial PR for a brand that is keen to be progressive. Ethical considerations are not just a matter of reputation and risk.

When employees and customers engage in trusted relationships, the rewards look set to be significant. A study by PricewaterhouseCoopers (PwC) shows that AI is becoming so important, so powerful, that the firm forecasts an upswing in global GDP of up to 14% (the equivalent of US $15.7 trillion) by 2030 as a result of the accelerating development and adoption AI.[4] The McKinsey Global Institute expects that around 70% of companies will adopt at least one type of AI technology by 2030, while less than half of large companies would deploy the full range.[5]

Yet the benefits might not be equal. Firms that can demonstrate ethical use of AI are slated to be more commercially successful. In a recent study looking at why the ethical use of AI is fundamental in attaining people's trust, Capgemini Research Institute[6] found the following results amongst the consumers it consulted.

62% **would place higher trust in a company whose AI interactions they perceived as ethical**

61% **would share positive experiences with friends and family**

59% **would have higher loyalty to the company**

55% **would purchase more products, provide high ratings and positive feedback on social media**

As these stats illustrate, trust in the ethical use of AI is incredibly important to consumers and they are willing to reward brands they view as such with their continued business.

[4] The macroeconomic effect of artificial intelligence, PricewaterhouseCoopers, 2018
[5] Notes from the AI frontier: modelling the impact of AI on the world economy, McKinsey & Company, Sept 2018
[6] Why addressing ethical questions in AI will benefit organisations, CapGemini, 2019

THE ADVENT OF REGULATION

Governments around the world are also stepping into the arena. A number of them have adopted guidelines and principles that aim to guide organizations in designing and implementing ethical AI.

Whilst these frameworks are voluntary today, governments are discussing mandatory legal requirements for AI. The European Union has announced its intention to regulate certain aspects of AI development and use, which may instigate a spiral effect of AI legislation globally.

Given data is the backbone of AI, many existing and upcoming laws are shaping the AI ecosystem. The European Union's General Data Protection Regulation (GDPR) that governs the use of personal data also applies to data driven solutions such as AI. Article 22 of the regulation includes a 'right to explanation', so-called because organizations must be able to provide 'meaningful information about the logic involved' in automated decisions. In parallel, the European Union is accelerating the trend towards a more regulated data market with new laws on data access and data sharing as well as online digital services.

In particular, automated decision-making is considered by many governments as an AI use that merits regulatory intervention. For instance, Canada has published a Directive on Automated Decision-Making. The Directive, a key pillar of the country's commitment to ethical AI practices, centers around the Algorithmic Impact Assessment (AIA), a tool that determines exactly what kind of human intervention, peer review, monitoring, and contingency planning any AI tools designed to serve citizens will need. The US Federal Government has also addressed the issue in proposed legislation around automated-decision systems. Similarly, the Australian Human Rights Commission is conducting wide consultations on AI-informed decision making.

The use of certain AI applications, such as facial recognition and remote biometric identification systems more generally, causes concerns for many governments that would like to limit potential adverse effects on individuals' rights and freedoms through targeted regulation. For example, in the state of California, facial recognition technology has been banned from use in public sector applications in some cities.

The AI regulatory landscape is evolving rapidly. Although most governments agree on the ethical principles that should be enshrined in AI development and use, we can expect diversity in the nature and stringency of future regulatory intervention in countries across the world.

PUTTING ETHICS INTO ACTION

This brief overview of the evolving ethics landscape highlights the growing seriousness with which consumers, citizens, governments and industry bodies are taking AI ethics. Yet, while discussion of ethics can often feel very esoteric, there are practical steps your organization can take to put ethics into action.

DEFINING ORGANIZATIONAL GUARDRAILS (AKA PRINCIPLES)

Early discussions of ethics in AI focused simply on architecting AI's FATE: ensuring the fairness, accountability, transparency and explainability of the solutions. Today, a number of public and private partnerships have defined an increasingly expansive set of frameworks to guide the ethical use of data and AI. Some of them include The Future of Life Institute's Asilomar AI Principles and the World Economic Forum. The EU has also developed principles for ethical and trustworthy AI, as has the IEEE and the OECD.

These principles and frameworks aim to inform broad, generally applicable standards as well as guiding development of applicable regulatory frameworks. They also support individual organizations in defining and instituting their own AI principles and operating boundaries. Certainly, the business domain, problems being solved, relationships, geographic and cultural identities—just to name a few—all inform an individual organization's priorities and ethos. However, these readily available resources negate the need for organizations to start from scratch when defining their own operational guardrails.

THE IEEE GLOBAL INITIATIVE ON ETHICS OF AUTONOMOUS AND INTELLIGENT SYSTEMS (A/IS)

GENERAL PRINCIPLES OF 'ETHICALLY ALIGNED DESIGN' ARE AS FOLLOWS:

HUMAN RIGHTS

A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.

WELL-BEING

A/IS creators shall adopt increased human well-being as a primary success criterion for development.

DATA AGENCY

A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.

EFFECTIVENESS

A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.

TRANSPARENCY

The basis of a particular A/IS decision should always be discoverable.

ACCOUNTABILITY

A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.

AWARENESS OF MISUSE

A/IS creators shall guard against all potential misuses and risks of A/IS in operation.

COMPETENCE

A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

INSTITUTING GOVERNANCE

Just as insight without action is pointless, so are principles that are not put into practice. Done right, governance establishes accountability for and transparency into AI-driven outputs and practices.

All software, including AI, is designed to deliver a defined outcome—a decision, a recommendation, an action. The question is: who decides what the outcome should be, and what steps are required to ensure the desired objective(s) are achieved? The very nature of the decision-making process shapes the end result because those involved in its inception bring all their knowledge, preferences, prejudices and unconscious experiences to bear. Diverse perspectives are therefore required to bring to light and then mitigate unconscious blind spots. Of course, once you have arrived at a use case, discussions of fair and appropriate use are not over. Just because AI decisions are machine based and not subject to all the same cognitive failings as human decision-makers, they are not infallible. AI systems can and do make mistakes.

Transparency is vital because it allows organizations to make considered decisions regarding risk and reward. Governing the AI development process from concept to deployment helps identify and mitigate a spectrum of risks. This includes the propensity of poorly conceived or architected AI solutions to foster inequality or accentuate bias. Thoughtful analysis and validation throughout the process can help uncover such issues even in the absence of fully explainable AI systems.

Emerging governance practices include management review boards to vet proposed applications, implementing model development standards that incorporate frequent checkpoints with diverse stakeholders, routine monitoring and review of results and outcomes, communicating where AI is being deployed and providing recourse for those impacted to understand and/or appeal decisions made by automated systems.

LEVERAGING EMERGING TECHNOLOGY

With AI, there is often a balance between accuracy and explainability: more complex models may be more accurate but are also harder to understand and explain. Determining the right balance requires case-by-case analysis. In some cases, the inability to explain the inner workings of a model may negate its use—particularly in highly regulated or high-risk sectors such as finance or healthcare.

Therefore, increasing model explainability (sometimes referred to as explainable AI or XAI) and interpretability is an area of active development and usable techniques exist today. Explainability focuses on describing the logic or factors by which an AI algorithm reaches a given conclusion or result. Interpretability focuses on mechanisms—such as the use of natural language generation—to explain how models work and to report their results in non-technical terms.

In addition, well known data profiling and visualization techniques can be used to validate the integrity and completeness of data input to an AI algorithm as well as the outputs. Emerging approaches to improving data privacy (such as differential privacy which helps anonymize discrete records without losing realistic data representations), the ability to utilize smaller data sets and synthetic data creation (to ensure target populations/conditions are adequately represented in both training and testing data) are also key enablers for AI.

Technology cannot, in and of itself, create ethical systems or make ethical judgments. But, just as emerging AI technologies have the potential to accentuate poor decision making, technology be applied to help mitigate these risks.

FINAL THOUGHTS

The disruptive, life changing power of AI is very clear to see. As adoption of this technology accelerates and is embedded into our lives in increasingly impactful ways, consumers, citizens, governments, regulators and interest groups are becoming increasingly concerned about its fair and equitable use.

While we all want AI to be employed in ways that do good in the world, ethical challenges are inevitably being exacerbated by the sheer scale at which it can be deployed, and the incredible range of applications it has. The situation is further complicated by the highly complex nature of modern algorithms—being probabilistic, data-driven and whose internal logic is often extraordinarily difficult to unravel, even for seasoned data scientists. It is this black-box or opaque nature of AI that creates trust issues, yet conversely, when AI is leveraged appropriately, it can deliver huge value to organizations and people.

As a leader, this conundrum of ethical challenges can perhaps feel rather daunting and any discussions around ethics can seem esoteric and rather removed from the practicalities of business and the focus on competitive advantage. However, as we have seen, there are some very good reasons to act now and ensure that your organisation engages in the ethics conversation.

Firstly, consumers demand the ethical use of AI, and studies show their intention to withdraw from brands that cannot demonstrate a concern for using AI ethically. Regulators are also increasingly focused on this issue and are likely to continue to be so. Finally, ethical use of AI is central to a modern, forward-looking corporate social responsibility strategy.

Fortunately, you will not be alone as you find the best ways to apply AI in your operating environment. A good deal of work has already been done by early adopters, collectives and interest groups to guide your way ahead. It's time to take the first practical steps in defining what feels ethical for your organization and implementing your own set of principles, and transparent, governed processes—doing so will help you to capitalize on the incredible potential of AI in ways that build trust with your employees, your customers or the communities and stakeholders you serve.

Find out more about how SAS can support you and your organization in building confidence in AI »

SAS.COM/EXPLOREAI

The content of this document is written and provided for general information purposes only. The document is not intended as advice on legal or regulatory matters, and should not be used as a substitute for taking specific advice on any of the topics covered.