Microsoft

Joint insurance industry point of view paper:

We need rules of the road for responsible AI and data science

Jim DeMarco, Senior Industry Digital Strategist, Microsoft • **Gregory Falco**, Assistant Research Professor, Johns Hopkins University • **Jerry Gupta**, Senior Vice President, Swiss Re • **Jonathan Silverman**, Industry Executive, Microsoft

Table of Contents

The opportunities and risks of artificial intelligence (AI)

Every major technological and commercial disruption in human history has relied on some form of insurance to help people safely adopt change. With the creation of cyber-physical hybrid technologies and the use of artificial intelligence to guide and even make decisions, the insurance industry can again blaze a trail of trustworthy innovation.

This white paper explores the need for responsible AI, as well as the new risks that AI poses. It further investigates the role of the data scientist who creates decision insight in a world influenced by AI. In this paper, we argue that data scientists will play an outsized role in AI-enhanced decision making and therefore should be governed by a code of conduct regarding how they create decision insight.

This paper also touches on how the insurance industry, as a leader in driving the adoption of new technology, has a responsibility to lead in the crafting of such a code of conduct, starting with our own data scientists and our own use of AI. The same challenges that apply to insurers also apply to their corporate customers who use AI. Insurers consume more data than most other industries, which implies not only the need for insurers to mitigate the risks of inappropriate use of that data in insurance but also the responsibility to set standards that can be applied both within and beyond the insurance industry itself. Accordingly, we propose a set of core principles and guidelines that would make up a code, the proverbial rules of the road in responsible AI, that should be applied by insurers and their customers.

We intend for this paper to drive a conversation around ethics in AI and data science that brings the insurance industry and our partners in social policy, as well as the businesses we insure, to a common point of view on how we should use AI and how we should do our job as data and decision scientists. To that end, we very much welcome feedback on this paper and wish to spark broader conversation across the data science community. See the conclusion for contact information.

The Fourth Industrial Revolution is underway, marked by massive social upheaval, economic disruption, digital transformation and, just like with previous industrial revolutions, seemingly unlimited opportunity. Perhaps the greatest driver of change in this age is artificial intelligence (AI). AI is showing great promise for society, from embedded AI used to speed disease diagnoses and accelerate time to market for new vaccines, to AI-infused communications that enable people to talk with each other in their native tongue despite not knowing each other's language, to optimized traffic routing that reduces the carbon footprint of shipping goods between continents. Digital transformation has fueled the acceleration of AI adoption, which helps businesses to be more competitive. But AI has also given rise to ethical and privacy risk, with bots driving divisive social media attacks, automated identity theft, and a host of other challenges previously unheard of. Perhaps more ominously, AI poses a new challenge in one core area that no other technology has yet touched: human decision making.

The Fourth Industrial Revolution differs from the three previous ones in that, for the first time, we can rely on technology itself to make decisions. Done right, AI can enhance human judgment by identifying and removing known mistakes. Done wrong, AI can make or exacerbate mistakes and guide the humans who make decisions into egregious or even evil behavior. The AI practitioners we call *data scientists* have the burden of defining this new technology in a manner consistent with global moral norms. Data scientists, and the applications they build, must be guided by a set of well-defined ethical principles.

At its core, AI is fairly simple. AI uses machines to mimic human behaviors to make decisions and interact with us based on massive amounts of data that no one person could remember or comprehend. Any AI application has three main contributors: the decision algorithm, the data that goes into the algorithm, and the algorithm's author (the data scientist). To get better at its decisions, AI requires increasing amounts of data. The role of the data scientist in AI is to select the data and apply the rules that define the AI algorithm.

The data scientist's success, and by extension the AI algorithm's success, relies on the proper application of both rules and data, which creates a conundrum: how does a data scientist get good data, and what rules should the data scientist apply to create the right algorithm?

The size and scope of available data

To drive accuracy, AI typically requires increasing amounts of data to validate and correct algorithmic output: the more specific the data, the more accurate the result, usually. The quality, accuracy, and completeness of data impact the effectiveness of an algorithm. When determining the scope of your data, it should accurately represent the specific populations that would be affected by the algorithm. Incomplete, inaccurate, or biased data can result in adverse or skewed results. Personal data requires compliance with applicable privacy laws and regulations.

AI's need for voluminous data runs counter to legal privacy requirements for data minimization and potentially challenges the human right to privacy. The scope of data availability for AI may be further limited due to the legal requirement for individuals' explicit consent before processing their personal data. Individuals should have control over the use of their personal data and the decisions made using it. Additionally, privacy laws require that adequate data protection controls be in place when processing sensitive personal data to reduce the risk of discrimination. This includes data revealing racial and ethnic origin, religious beliefs, gender, or any personal attribute that could contribute to discrimination against an individual. This use of data necessitates the need for ethics, accountability, and compliance with legal requirements when considering the size and scope of the data.

The problems of discrimination in AI

Insurance companies using AI to improve their ability to pool and mitigate risk can expand coverage to individuals who may not have insurance or may be underinsured. However, there are concerns that the increased use of AI may exclude or discriminate against specific groups inequitably. Marginalized or vulnerable groups may be susceptible to discrimination due to biases embedded in data attributes such as income, race, or gender. Discrimination also occurs when a proxy is used that is strongly correlated to these data attributes (for example, using employment history or access to credit, which may be tied to race or gender).

Discrimination in AI can have devastating consequences, such as:

- Imposing financial disadvantages for low-income families that can't qualify for affordable financial products.
- Improperly defining risks, leading to forcing inappropriate or unsuitable products on customers.
- Perpetuating systemic poverty within a community by precluding financial support and freedom.

Discrimination could be caused by bias built into AI at various stages. It can occur due to misinterpreted data, bad data, errors in machine learning (ML), false assumptions, or false conclusions. Discrimination is challenging to prevent because it can't be found easily until it is identified in the output. Here are some examples of how bias in AI can result in discrimination:

Data collection: Data should accurately represent the specific populations that would be affected by the algorithm's output. Machine learning is only as good as the data that is used for training, so "bad data" (incomplete, inaccurate, or biased data) can result in a discriminatory output. For example, a car insurance company using driving records to determine premiums could be using biased data based on tickets that are disproportionately issued based on race. Discrimination can occur even with good (accurate and unbiased) data if a proxy is used that is correlated with a protected attribute such as race or gender, causing indirect discrimination.

Algorithms: Discrimination can be built into the algorithm. An algorithm used in one circumstance may cause discrimination when used in another similar circumstance. When the desired output of the algorithm is to maximize profits or gain efficiency, there may be little focus on fairness. Also, weight may be placed on data attributes that could inadvertently cause bias in the algorithm. For example, putting weight on zip code for property insurance may adversely impact low-income communities.

Model training: AI relies on learning human behaviors and practices to mimic human decision making. Human behaviors are not free from bias and errors. Bias can be taught during the machine learning process based on mimicking discriminating or biased human behaviors. For example, a company's hiring model based on historical hiring decisions was determined to be discriminating against women based on learned human hiring decisions.

Malicious intent: Moral problems related to data use may intentionally build bias into an algorithm to derive a desired output. For example, companies redlining specific communities with lower property values make it more difficult for individuals in that community to get loans to purchase homes and establish businesses.

With the massive amount of data processed in AI, even a minor error causing discrimination can potentially have far-reaching, detrimental consequences.

To avoid discrimination, we should prioritize fairness and transparency at every stage of AI. Those involved in AI must develop ways to predict, prevent, and monitor bias. Additionally, input from a diverse group brings different perspectives and viewpoints to managing bias. Algorithms should be scrutinized by diverse teams from different functions, using multiple tests to validate that the algorithm and its output are in line with the expectations of the business and regulators. Because discrimination or bias can exist at any stage of AI, it is important to raise awareness and provide training.

The three main contributors to artificial intelligence

At its core, AI emulates the decision making that humans do naturally. To emulate this decision making, AI uses two key inputs: a decision-making algorithm that establishes the rules machines use for the decision, and the data that guides how the algorithm is applied. Beneath those two inputs are the choices made by the data scientist as to how the algorithm is formed, what data is used for input, and how the algorithm interprets the data. Thus, there are three major contributors to AI in real life: the data, the algorithm, and the data scientist who crafts and orchestrates them. We'll consider each of these contributors in turn.

The data

AI works on translating information (data) into patterns to guide decisions. Before that data can be processed by an AI algorithm, it often requires cleansing because the algorithm may not be able to interpret the data in its purest, raw form. Data collected from sensors needs to be cleansed because there may be a wide degree of variation caused by anything from human error to a faulty device. Data cleansing generally involves selecting boundaries for outliers that need to be omitted so that trends are more easily interpretable or refining sample sets to target particular decision parameters.

Because data is subject to these lifecycle processes, two individuals with the same raw dataset and the same algorithm could arrive at different results due to cleansing decisions. Therefore, transparency about what is done to the data before it ever meets the AI algorithm is critical to the ethical application of AI.

The decision algorithm

AI algorithms are generally built as a set of steps to interpret patterns in data and make inferences about the value of different decision options about those inferences. The earliest AI algorithms were generally driven by simple, traceable rules engines. As time, underlying technology, and computational capacity have progressed, the power and ability of these algorithms have grown tremendously, yielding far more predictive insight from representative data. At the same time, the complexity of these algorithms has increased, often reaching beyond the ability to draw simple references back to how any given decision was made. Further, as more data is input over time, especially when the algorithm is deployed in a real-world context, the refinement of the algorithm's inferences can yield both unexpected and untraceable conclusion instead of inferences for the second time. Such inferences are not themselves good or bad, but their application in decision making can yield morally indefensible actions in the wrong circumstances.

Algorithms are quickly moving from the domain of decision support tools to that of social constructs, similar to the concept of property ownership or the creation of legal entities. Like any social construct, data algorithms now require a social contract, supported by regulations. Bogost[1] warns about the impending "computational theocracy," which is evidenced by the Centrelink scandal in the UK, where incorrect computer-generated debt notices were sent based on opaque algorithms. Algorithms create a false perception of unbiased decision making. This in turn risks creating a vicious cycle where algorithm-driven processes flag the disadvantaged and reinforce those algorithms with each new flag, thereby creating a permanent class of the economically poor and permanently eroding social mobility.

Further, not all AI algorithms are created equal. The use cases for the different types of AI algorithms and the requirements for the algorithms vary drastically. Two questions will be critical to future ethics considerations: whether the right type of algorithm was selected for the problem at hand, and how the algorithm is curated over time in real-world use.

The data scientist

Data science is a merging and blending of many quantitative fields, such as computer science, actuarial science, financial mathematics, statistics, operations research, and econometrics. Each of these core theoretical fields has individual best practices and algorithms of choice, and blending them brought a rush of new approaches to the forefront. The emergence of data science, mixed with the advancement of the graphics processing unit (GPU), has made it possible to run mathematical algorithms on an individual PC. Data scientists deal with increasingly complex, opaque algorithms, and advanced approaches to quantitative science. Because their work can be so far-reaching

[1] Ian Bogost, "The Cathedral of Computation," *The Atlantic*, January 15, 2015.

and so hard to explain, data scientists must develop tools to make sure their work clearly yields a business-appropriate result that does not unfairly impact the people affected by their work.

Because they use and manage tools to influence or even make decisions that were once the province of human decision makers, data scientists bear some of the burden of ensuring that their work is properly informed by norms that would otherwise apply to those decision makers. In insurance, for example, the primary decision maker around risk is the actuary. The work of a data scientist in support of actuarial decision making, therefore, must be subject to the norms governing actuaries.

Why actuaries have a code of conduct

Actuaries must make decisions that could have particularly negative effects if made incorrectly. Their job is to be the stewards of financial well-being, not only for the insurance customers but also for their insurance agencies. Often balancing conflicting interests, actuaries are tasked with properly assessing risk so that they can provide the financial stability gained through insurance to as many people as possible while also keeping their insurance companies solvent.

An actuary's decision to provide insurance, or provide overly expensive insurance, plays a significant role in an individual's ability to maintain a financial safety net. So, it is imperative that these decisions are made using appropriate methodology and that the risk is accurate and not unfairly discriminatory. The weight of these decisions requires a thoughtfulness that is imbued within the actuarial code of conduct to which all actuaries must adhere.[2]

The emerging responsibility of the data scientist

Actuaries are already familiar with the great responsibility they bear. As data scientists see an increasing influence over actuarial decision making, it's imperative for them to share in the responsibility for their influence. Data scientists need to extend their professional responsibilities beyond merely creating algorithms with accurate results to ensure safe and valuable ones as well. These expanded responsibilities include multiple facets:

- **Building:** Ensuring that valuable insights are gained from the data and lead to actionable decisions (this is a performance metric and has always been a responsibility).
- **Sourcing:** Ensuring that the data is ethically sourced and that any bias or incompleteness in the data is understood and accounted for in modeling.
- **Architecting:** Ensuring that the algorithm chosen, including the judicious selection of metrics for optimization, provides valuable insights that are related to the problem. This architecting includes understanding the limitations of the algorithm and its behavior in cases of extrapolation.
- **Business translation:** Explaining and providing knowledge transfer to the end users or group whose decisions are being automated, or whose decisions are based on the insights generated. Algorithm users need to understand the limits of the algorithm, the confidence levels it generates, and when to use it or not use it.

[2] *Code of Professional Conduct*, American Academy of Actuaries, https://www.actuary.org/content/code-professional-conduct.

The need for rules of the road

The adverse consequences of biased algorithms can be as life-altering as those of legal or medical malpractice:

- **Socio-economic:** Consumer finance and insurance decisions are increasingly being driven by AI systems. Everything from your ability to buy a house or car to your credit limit is controlled by AI. Once a consumer gets caught in the AI trap, it can become very difficult to get out of it. Digital offerings are heavily influenced by AI-driven personalization techniques, which can play a big role in limiting consumer choice.
- **Political-regulatory:** Data-driven policing and discriminatory policymaking can result from funding and policy decisions made as a result of bad data or biased algorithms trained on data that favors homogeneity.
- **Public health:** Biased algorithms could lead to the very real possibility of rationing care across the public based on inappropriate characteristics, yielding healthcare allocation that is both discriminatory and ineffective.
- **Excessive trust in machine results:** The air of authenticity lent to a decision because "numbers don't lie" in computational models can result in an overreliance on the outputs of AI algorithms. Most data scientists rely on the computational model of Design, Measure, Analyze (DMA), which has the perception of being a scientific methodology to root out mistakes; however, the variable means by which DMA is applied make it more of an art than a science when ensuring proper outputs. As a result, DMA has created a social construct that protects the perceived integrity of the outputs even if the data that feeds AI is riddled with existing biases, exclusions, and noise and the algorithmic output strays from social norms.

So, because of the risk of adverse consequences, just as with the legal and medical profession, data scientists need sustainable governing norms—rules of the AI road.

The responsible application of AI: rules of the road

AI requires rules to guide machines' decision-making process in its two primary applications:

- Computational AI: the use of machines to process information to make decisions.
- Cognitive AI: the use of machines to emulate human actions to generate information for decision making.

Guiding the data: ethical sourcing and open data

Whether computational or cognitive AI is applied, the AI itself is built on, and grows because of, the ingestion of data. When applied in insurance, data drives analytic, actuarial, and underwriting decisions. An insurer evaluates an applicant by comparing them with an existing data set to determine their risk. So, the data used for comparison purposes needs to be sourced in a manner that's appropriate and from materials that are unbiased.

Mortgage underwriting is a classic example of the problems with biased data. It uses historical mortgage data based on decades of redlining aimed at preventing Black loan applicants and other people of color from buying homes in white neighborhoods. Using this biased data perpetuates the effects of racism. Our goal is the ethical sourcing of unbiased data to lead to more equitable decision making.

Initiatives like the Open Data Initiative, founded and developed jointly by SAP, Adobe, and Microsoft, aim to combine ethically sourced and unbiased data together transparently into a single data lake that can be used to derive AI insights.

Guiding the algorithm: responsible AI

AI's two main applications—computational and cognitive services—rely on a machine to gather data inputs, process them according to a decision-making algorithm, and provide outputs. Cognitive AI's outputs are frequently used as inputs to computational AI's decision-making engine. What makes AI different from other forms of computing is that AI algorithms are *self-learning*; that is, the outputs of an AI algorithm's computation are reused as inputs in order to provide inferences that guide how the algorithm will make future decisions. Therefore, the initial algorithm (the machine learning model) must be built in a way that is not only accurate to some degree out of the gate but also adaptable, to remain useful as it changes according to learning inferences. In other words, for AI to be usable, the decision-making algorithm must provide a trustworthy result,[3] not only in the initial application but also over time.

In 2018's *The Future Computed: Arti icial Intelligence and Its Role in Society*,[4] Microsoft proposed that designing AI that remains trustworthy over time requires creating solutions based on ethical principles "deeply rooted in important and timeless values." Microsoft identified six core principles to guide responsible algorithmic development: fairness; reliability and safety; privacy and security; inclusiveness; transparency; and accountability.

Microsoft identified six core principles to guide ethical algorithmic development:

- Fairness
- Reliability and Safety
- Privacy and Security
- Inclusiveness
- Transparency
- Accountability

[3] Note that trustworthy does not necessarily mean predictable. AI algorithms can provide random results for a given decision by design in order to provide a more accurate result on an aggregated set of decisions. Further, because AI algorithms are self-learning, output results from one AI computation to the next are expected to vary as the machine gains inferences.

[4] For a detailed description and a free downloadable copy of the book, see https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/

Because AI algorithms self-learn, building a trustworthy algorithm requires addressing ethical principles at each step in the algorithmic lifecycle, from initial envisioning through algorithmic definitions, prototyping, initial launch, application, and assessment for evolution and re-envisioning. The six ethical principles must be adhered to at every stage in the life of an AI algorithm, yielding a Responsible AI Lifecycle (RAIL):

Fairness

AI systems should treat all people fairly, applying no biases. Indeed, if properly deployed, AI should detect and eliminate bias in most decision algorithms.

Safety and reliability

Safety and reliability must be considered not only in purpose-built circumstances but also in unexpected conditions, including when systems are under attack. AI systems must be tested extensively, updated based on human user feedback, and monitored for ongoing performance.

Inclusiveness

Everyone should benefit from intelligent technology. Like fairness, AI systems should empower everyone and engage people regardless of age, gender, race, or physical/mental capabilities. This tenet is fundamental for a sustainable AI system.

Privacy and security

AI is driven by data, which must be secured against external influence and tampering.

Data privacy and security must be more than mere regulatory compliance considerations that need to be incorporated in all aspects of the insurance lifecycle. Privacy and security must be treated as core customer rights if the insurance or technology industry is to provide AI solutions that people trust.

Transparency

When AI systems are used to help inform decisions that have a tremendous impact on people's lives, it is critical that people understand how those decisions are made.

Transparency in AI is about understanding the steps taken and the final reason behind how a machine decides. In part, transparency means that those who build and use AI systems should be forthcoming about when, why, and how they choose to build and deploy their systems, as well as their systems' functionality. Transparency also means that people should be able to understand and monitor the technical behavior of AI systems.

Accountability

The people who design and deploy AI systems must be responsible and accountable for how their systems operate. Accountability helps ensure that AI systems are not the final authority on any decision that impacts people's lives and that humans maintain meaningful control over otherwise highly autonomous AI systems, especially when AI systems make consequential decisions.

To ensure that people remain ultimately accountable for AI systems and their operation, those who manage AI systems on a daily basis should be trained to understand the design and limitations of the AI system, and they should have the authority to remediate as necessary. Organizations should also consider establishing a dedicated internal review body to guide practices regarding the development and deployment of AI systems.

Governance

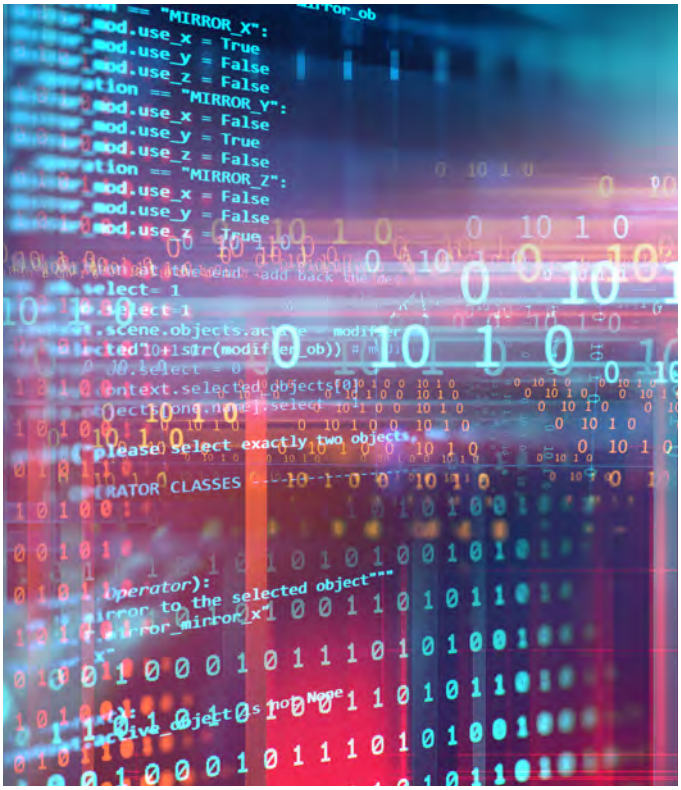
These six principles determine *whether* a particular AI algorithm can be deemed ethical. Next we will present a framework for how to ensure these principles are properly applied.

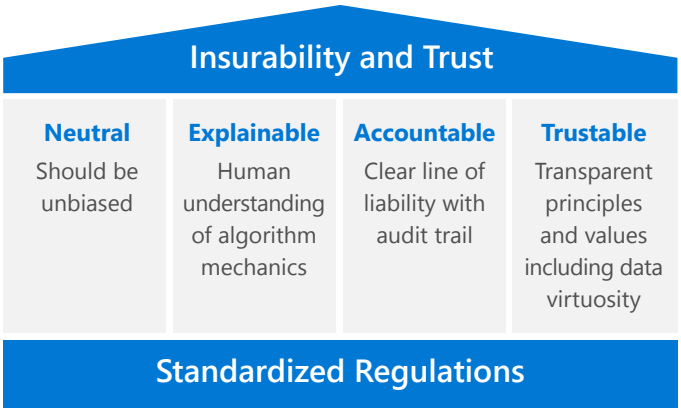
Creating an actionable framework requires implementing a smaller set of practical guidelines for determining whether any given system meets the standard laid out in these principles. This framework derives a four-part test for applying the principles, in some cases repeating the core principles or combining principles to make their application more readily understood.

Guiding the designer: the **NEAT**[5] Framework for data science practitioners

While AI is powerful and can benefit mankind, one has to be aware of its flaws and the impact it can have:

- There is an intrinsic lack of explainability in some of the more sophisticated algorithms, which can result in inaccurate or biased outputs that would be difficult to identify and cure.
- One of the larger risks of AI is societal. AI-based systems are making life-altering decisions in the space of financial services, medicine, university admissions, and human resources, for example. The potential for bias based on race, gender, age, and ethnicity is high, while mechanisms to prevent this bias are not strong.
- In the case of autonomous systems, it would be difficult to assess liability. This assessment is especially problematic given the movement to make AI systems "legal persons."
- Malignant actors and governments face the temptation to use AI systems in ways that infringe on people's privacy for personal gain or control.
- All software has bugs. However, in self-learning AI systems, it may be more difficult to detect and fix the bugs.





To address these issues, we propose the NEAT Framework for algorithm development. We believe that algorithms should be:

- **N**eutral: Algorithms should be neutral with no unintentional bias. Certain functions, such as credit card approval, have an inherent (and accepted) bias based on income. However, these biases should be stated upfront when developing algorithms. Furthermore, tests should be conducted to ensure that social bias is not present in the algorithms.
- **E**xplainable: We propose that algorithms should be rated based on risk to society and that any algorithm that has a societal/human impact must be explainable. For instance, an image recognition algorithm may not require as much explainability as an algorithm that is making mortgage underwriting decisions.
- **A**ccountable: There has been talk of making algorithms "legal persons" for the purposes of the law. We strongly advocate that there be a clear line of liability associated with each algorithm. Just as a doctor making an error is liable, so should a data scientist making an error in algorithms that impact society/humans be liable. It is important that a natural person be responsible to ensure that there is accountability and recourse. Versioning and commenting standards should be developed, along with best practices for logging, to ensure that a foolproof audit trail exists.
- **T**rustable: The developer and approver of the algorithm should clearly disclose the purpose, principles, and values that were adhered to in developing the algorithm. This self-attestation ensures that the people involved in building the algorithm are purposeful in the actions they take and are mindful of the impact and consequences.

[5] ©2017 Jerry Gupta

Calling for a code of conduct for data science

As ethics continues to emerge as the cornerstone of trustworthy artificial intelligence, there is a need for a meeting of the minds concerning the responsible creation and use of algorithms. This consensus could materialize as a code of data science conduct. As with any code, be it a building code or code of ethics, a code of data science conduct will be largely self-imposed and reported at first. The code can serve as a baseline set of guidelines that helps aid decision making when artificial intelligence algorithms are developed. Such a decision-making tool can be pointed to in the case of an audit or even for purposes of explaining the boundaries of the algorithm's functionality.

Codes evolve over time. They are not developed in a vacuum, but through a collective decision-making process that is filled with heated debate and frustration—and ultimately, a beneficial exercise. While developing a code of data science conduct, there must be an arbitrator who shepherds the discussion. The arbitrator should be impartial and not necessarily benefit from the shape of the final code—instead, their motivation and benefit should be tied to the fact that a code of data science conduct is ultimately developed. The insurance industry is uniquely positioned to be sufficiently motivated and equally impartial.

We believe that insurance should be out front advocating for data science ethics because it is the best candidate for the task:

- **Insurance is one of the heaviest users of data and complex algorithms.** Insurance uses more data than many other industries in the global marketplace. It is built on analyzing large amounts of data from disparate sources to identify and measure risk and make life-changing decisions. The volume of data collected and analyzed gives insurance a great deal of experience and insight into data use and increased incentive for establishing standards to ensure it is used responsibly.

- **Insurance is the best assessor of risk.** Insurance is built on assessing risk using analytics, which gives it the authority and expertise to identify the risks of using problematic data and algorithms in AI.

- **Insurance has the ability to enforce adoption.** By withholding business owner insurance or directors and officers (D&O) insurance from companies that don't follow the agreed-on code of conduct for AI, insurance providers can encourage faster and wider adoption by creating a financial incentive. Leaders who would be exposed to personal liability without these insurance policies in place would then have a reason to push their organizations for change. We believe this type of enforcement would be more effective than government mandates.

While these rules of the road should apply to every industry that uses AI, insurance is uniquely suited to understand, drive, and guide change. Insurance has a mechanism and an interest in governing the responsible application of AI. Accordingly, following its traditional role in innovation, insurance needs to take the lead.

We welcome all participants into the conversation on this important topic! If you have questions, comments, or other feedback on this paper, or would like to join us in advocating for and formulating an industry-driven set of guidelines for responsible AI, please contact insaiethics@microsoft.com.