# Industry needs to embrace data ethics: here's how it could be done

AUTHORS (listed in alphabetical order)
Bethan Cantrell, Senior Privacy Strategist, Business & Corporate Responsibility, Microsoft
Javier Salido, Principal Program Manager, Business & Corporate Responsibility, Microsoft
Mark Van Hollebeke, Privacy Practitioner-in-Residence, Data & Society Research Institute

## ABSTRACT

IT industry companies should willingly embrace data ethics and the increased transparency and enhanced accountability that ethical data use will require. Microsoft is actively engaged in discussions with other leading IT companies about why and how to adopt some basic industry norms for developing advanced analytics, especially AI. This paper opens by presenting compelling reasons for adopting ethical design practices relevant to the viability of data science as a business practice. It then proposes some principles and values related to data ethics, how these might become operationalized norms, and explores what ethical issues would be addressed by doing this. While the integration of ethics into business practices will have benefit to each business that adopts it, large-scale societal benefit will result only if data ethics are adopted industry-wide.

## WHY EMBRACE DATA ETHICS?

Privacy advocates, academics, and others may fear that quarterly profit reporting and general economic pressures create systemic resistance to ethical practices, which can only be overcome by regulatory pressure. Regulation may in fact, be needed. However, there are compelling business reasons that are inspiring Microsoft and other IT companies to initiate internal and external discussions on how to manage the increasing ethical complexity of data science, especially AI and deep learning.

In fact, industry sustainability is at the heart of this work. Simply put, the harm resulting from negligent design will likely dramatically outweigh the overhead of operationalizing principled design practices and ethical review of advanced data innovations.

There are compelling reasons for adopting principled design practices that are relevant to the viability of data analytics as a business practice. Such reasons include (but are not limited to) the following:

- Building and maintaining trust with customers
- Creating a viable and sustainable data ecosystem for use by advanced analytics tools
- Encouraging customer use of advanced data-analytics services
- Encouraging customer adoption of cloud services
- Operationalizing norms ahead of regulatory requirements
- Helping enterprise customers reduce risks in using advanced analytics
- Assisting partners and enterprise customers in staying ahead of regulatory requirements
- Creating a level and sustainable playing field across the IT industry concerning data analytics

Sustainable businesses, those looking far beyond the next quarterly report, eagerly pursue customer trust and positive brand image. Media and consumers rightfully lodge brand-damaging criticisms against businesses with unsavory or untrustworthy data practices. These impact usage and brand valuation. The recent history-making data breach at Yahoo is case and point. This incident puts the Verizon acquisition of Yahoo, or at least the terms of the deal, in question and further adds to a negative perception of Yahoo and media criticism of its current CEO.[i]

The future viability of data-analytics services is tied to trust. As these services increasingly sell business-intelligence and personal insights culled from customer data, enterprise customers and consumers need to see value returned from these analytics. Deriving valuable insights requires increased access to data for analytics, and increased access will be provided only when customers trust that good data protection and other responsible practices exist. Risk of exposure must be balanced with trust and demonstrated added value. This virtuous cycle, required for growth of the analytics and AI business, could easily become a vicious cycle of customers blocking access to data or purposively obfuscating or falsifying the data to which analytics services have access. Customers choosing to limit access to data, or purposefully falsifying or obfuscating data they provide, results in analytics with low value. It also undermines the

ability of the IT industry to reason over vast amounts of data and deliver value from it.

All IT companies should consider the business value of adopting principled design practices and ethical review. Adopting these data ethics practices should aim at preserving respect for individual agency and community self-determination, contributing to the improvement of societal equality, and reversing the growing power asymmetry between large institutions and individuals.

This view has helped frame Microsoft's vision for advanced analytics, especially AI. AI, from this perspective, is not about developing computer intelligence for its own sake. Rather, the aim is to augment human intelligence for the sake of empowering the full expression of human creativity. Satya Nadella, Microsoft's CEO, recently articulated this vision in various places including several developer conferences and his June 2016 essay entitled, "The Partnership of the Future."[ii] Microsoft is actively pursuing "what's possible when human and machine work together to solve society's greatest challenges like beating disease, ignorance, and poverty."[iii]

Thus, Microsoft is actively engaged in thinking through and constructing our approach to advanced data analytics. We are eager to collaborate with academics, advocacy groups, data science researchers and industry leaders concerning, as our CEO, Satya Nadella puts it, "the universal design principles and values that should guide our thinking, design, and development" of advanced data analytics technology, including AI.[iv]

## OPERATIONALIZING NORMS

Over the years, Microsoft has operationalized its values in a range of areas such as privacy, security, online safety and global readiness. We're well set-up to develop a set of practices for design-side ethical development and review processes.

Achieving this aim at Microsoft will involve operationalizing norms within our data science and engineering workforce. To begin with, this will include two interrelated efforts: 1) awareness and culture change programs that clearly articulate our values and how they impact our data innovation work and 2) crafting design-side processes to triage potential ethical impacts to particular data uses.

Microsoft's focus on using data to benefit our customers and society at large has been expressed in multiple places, from our new privacy principles to Satya's essay on AI, mentioned above.[v] In it, Satya proposes six AI "principles and goals, as an industry and a society, that we should discuss and debate" as well as four mandates for human beings.[vi]

The '10 Laws of AI,' as media outlets are calling them, are:

1. AI must be designed to assist humanity.
2. AI must be transparent.
3. AI must maximize efficiencies without destroying the dignity of people.
4. AI must be designed for intelligent privacy.
5. AI needs algorithmic accountability so humans can undo unintended harm.
6. AI must guard against bias.
7. It's critical for humans to have empathy.
8. It's critical for humans to have education.
9. The need for human creativity won't change.
10. A human has to be ultimately accountable for the outcome of a computer-generated diagnosis or decision.

While Satya's essay focuses on the societal impacts of the coming AI tidal wave, the principles he articulates there can be summarized in three core values: *respect*, *fairness*, and *harm avoidance*. These values speak to concerns that are being raised by regulators and other concerned observers. Spelling out and debating the merits of these values and the guiding principles they imply will be key.[vii]

These three values are complex. Fully expressed they include (but are not limited to) the following components:

<u>Respect</u>

- Always put people first
  - Empower people. Seek to increase agency and self-determination.
  - Create, rather than diminish, opportunity
  - Recognize human self-worth (dignity). Treat people as more than data points.
  - Respect human rights
- Be cognizant and respectful of social relationships and culture
  - Protect vulnerable populations
  - Empower cultural self-determination and independence
  - Seek to enhance the ways that people foster trust with each other
  - Allow for differences in social norms

Fairness

- Ensure the accuracy and integrity of data
- Aim to identify and understand the impacts of bias in data sets and mitigate that bias
- Protect against unjust impacts
- Ensure that the criteria we use to provide decisions about individuals and groups are rationally justifiable
- Be accountable for the insights and decision suggestions of our data-powered services

Harm Avoidance

- Seek to maximize individual and societal benefits while reducing risks of harm
- Where adverse impacts are unavoidable, seek to diminish those impacts (especially with consideration to vulnerable populations)
- Increase trust via transparency (e.g., by clearly communicating risks)
- Protect Microsoft's reputation and long-term business sustainability with positive outcomes
- Threat model against potential misuse, service reliability and impacts of failure rates

Tying into existing governance programs and leveraging the relationships and networks used by those programs will help ensure the operationalized values scale, and that our projects empower people, enable autonomy, create opportunity, and recognize individual self-worth. These values are already in use in our commitment to international human rights norms and the GNI principles, which require careful thought about the differences and nuances of various world cultures, including consideration of different cultures varying views of ethical behavior. Similarly, our global readiness program policies are also reflected here as these three values require thoughtful consideration for local cultures, applying appropriate social etiquette in our services, and consideration of how we impact social interactions.

The point is that by inculcating these values into the Microsoft culture, we simplify what employees do during design. Understanding these values and reflecting on them at design creates human-centric innovations that pave the way to better compliance and experiences downstream. Articulating and agreeing to norms is a first step, but of equal importance is operationalizing them within the data innovation design process.

## PROPOSED DATA ETHICS PRACTICES

For data science projects, we can construct a lightweight review process at design that integrates into existing business processes. Questions along the lines of those below enable a triage approach that can be used in projects to identify potential societal impacts and ethics concerns.

When designing the process and questions, it's important to remember that best practices on triaging ethical impacts will evolve as the industry advances, and the processes and questions should be updated accordingly. Also, the subject area and process specifics should reflect the nuances of each team's work.

> **Identify risks in the project**: Does the project attempt to determine anything about a person that may result, directly or indirectly, in harm to that person? Does it deal with a targeted or vulnerable population? Can the intended outcome reduce choice or limit future opportunity for anyone? If the answer to any of the questions is "yes," have domain specific subject-matter experts been consulted or involved directly in the project (e.g., demographers for use of census data, criminologists for recidivism studies, etc.)?
>
> **Identify risks in the output:** Will the result be shared externally or deployed into a production system, or is the output an intermediate (research) step that will help to further refine our understanding of the problem and challenges, and/or the quality of the data set? If the result may be seen as reflecting bias or the quality of the data is in doubt, could this result be used to unfairly target or discriminate against a population? Should the scope of use of the result be limited by hardening processing and restricting results, or contractual limits on liability, or other methods? Are there trusted subject-matter experts with appropriate domain specific knowledge on hand to assist in resolving any potential ethical issues?

Once an ethical concern is articulated and documented, employees may be able to mitigate risks by using data science itself to detect anomalies in the data, the product or the output. Microsoft researchers and other knowledgeable employees are pursuing such technical solutions. The 2011 Fairness through awareness article authored by Microsoft's Cynthia Dwork et al. has become

a foundational article on this topic, and more recent research and scholarship is very promising.[viii] There is *much* to learn in this area, and a robust and ethically-focused data culture would prioritize such research and testing among data scientists and researchers.

## THE PROBLEMS TO ADDRESS

At Microsoft, and presumably at most IT companies, there is a general sense that data is good, and more data is better! Most of our new products, services and features that use data innovations at Microsoft relate, in one way or another, to helping people make better decisions. Presumably we would all benefit if most decisions are made on the basis of data—information about the world. So how is it that ethical problems arise? Generally speaking, there are four aspects of data science that can lead to ethically thorny situations:

1. Bias in the data. Common assumptions in ML are that ''data is neutral," that it always represents "ground truth," and that if we have sufficient amounts of data and pay attention to it, valuable insights will emerge. Nevertheless, a dataset may inadvertently reflect the biases of the people or organizations that collected or curated it, and even those of society at large. We've learned, for example, that men are more likely to provide feedback on our products and so our feedback database may skew toward male views. And broad societal biases may explain the recent "three black teenagers" incident on Google search.[ix] This concern predates the world of big data. The seminal federal privacy legislation—the 1970 Fair Credit Reporting Act—was aimed squarely at reducing the risk of consumer harm from erroneous and perhaps biased data in credit reports.

2. Interpretability of algorithmic models. How algorithms inform a decision-making process is often opaque. A predictive analytics decision-making process that can be explained to a person in non-mathematical terms is deemed to be "interpretable." A system is "non-interpretable" when the underlying mathematics are too complex or abstract to explain in plain language, even to a data scientist. Data scientists often select algorithms based on how well they perform, without giving much thought to interpretability. This is of special concern when the project is one where there's potential for harm: the opaqueness of an algorithm can cause a negative perception if that same algorithm also denies credit or an employment opportunity. Interpretability is orthogonal to performance: a non-interpretable ML system can be very accurate in its predictions, but that opaqueness can also hide when the algorithm is poorly suited to the project.

3. Controversial uses of data science. Unexpected and controversial use of data, while primarily a privacy problem, can also have ethical implications. One of multiple recent examples is the Facebook emotional contagion study, where researchers filtered and otherwise manipulated users' individual news feeds to study how emotions expressed by others can influence a user's emotional state.[x] The study resulted in damaging news reports expressing alarm that Facebook might seek to manipulate its users' emotions and violate their expectation that their news feed is personal but objective.[xi] Facebook also received press attention recently for a patent filed in 2015 for a system that would enable lenders to assess a borrower's creditworthiness by checking their Facebook friend's credit scores. (People in vulnerable communities might fare poorly under such a system.)[xii] Once organizations have access to data, the desire to use that data is natural. Coupled with the argument that "data is neutral," review of underlying ethical considerations may be pushed aside. Direct benefit to the data subject, and the data subject's ability to understand the complexity of the data environment, must be considered with each project.

4. Unsupervised system decisions and interactions. Scenarios that used to be considered the domain of science fiction or philosophers (like the famous Trolley Car problem) are now being considered by regulators, and now must be considered by engineers. Typical examples are those of the autonomous vehicle that is about to crash and has to "decide" between running over a pedestrian or crashing into a wall and likely injuring its passengers,[xiii] and those involving the interaction of autonomous adversarial systems, illustrated by the flash-crash example of the futures and equities markets in May 2010.[xiv]

## CONCLUSION

At Microsoft, we feel that the first step is recognizing that these thorny issues exist; some data scientists are aware of them and others are not. For those who are aware there is a clamoring for guidance on how to think through these issues and design to avoid harm. We are currently at the stage of articulating guiding principles and triage processes teams can use, but there is a need to test them in practice and to vet and further develop them with our teams and with others in industry.

We also wish to ensure that we have the correct problem set and thus discussion and outreach across multiple stakeholders is important. Just as industry has learned to embrace privacy best practices in order to build consumer trust, it is time for industry to agree to basic data ethics norms that will extend that trust to the positive potential of big data analytics.

---

## NOTES

[i] Recent news stories note the impact of the Yahoo data breach on brand: "Verizon Puts Yahoo on Notice After Data Breach: Statement leaves door open for a potential renegotiation of Verizon's $4.8 billion acquisition of Yahoo," by Thomas Gryta and Deepa Seetharaman, updated Oct. 13, 2016, http://www.wsj.com/articles/verizon-sees-yahoo-data-breach-as-material-to-takeover-1476386718; "The bad news for Marissa Mayer somehow manages to get worse: A massive data breach affecting several hundred million users could impact the company's $4.8 billion sale to Verizon," by Maya Kosoff, September 22, 2016, http://www.vanityfair.com/news/2016/09/marissa-mayers-legacy-at-yahoo

[ii] Build 2016 Keynote by Satya Nadella, https://channel9.msdn.com/Events/Build/2016/KEY01#time=0m53s presentation and Ignite 2016: https://ignite.microsoft.com/#fbid=3PbswRfZzpp. "The Partnership of the Future," by Satya Nadella, Slate.com, June 28, 2016. http://www.slate.com/articles/technology/future_tense/2016/06/microsoft_ceo_satya_nadella_humans_and_a_i_can_work_together_to_solve_society.html.

[iii] "The Partnership of the Future," Nadella, Slate.com, June 28, 2016. See also the Microsoft roadmap to a trusted, responsible, and inclusive cloud, entitled "A Cloud for Global Good," that includes 78 policy recommendations for driving towards this technology revolution for all: https://news.microsoft.com/cloudforgood/, and http://news.microsoft.com/features/microsoft-ceo-and-president-team-up-for-oct-3-cloud-keynote/ , October 3, 2016.

[iv] Nadella, Slate.com, June 2016.

[v] *Ibid*.; See also the Microsoft privacy principles: https://privacy.microsoft.com/

[vi] *Ibid.,* Slate.com

[vii] It should be noted that this essay does not seek to provide the philosophical grounding for taking this approach or for selecting these particular values. Developed as a pragmatic middle ground between consequentialist and non-consequentialist ethical theories, they are meant to provide guidance for morally "good" results (enhanced well-being, increased opportunity, etc.). Because such results cannot be guaranteed, they also help designers act with integrity and do the "right" (and fair) thing wherever possible. These values also reflect previously articulated societal, academic, and industry norms, as well as the demands of regulatory frameworks. Rather than presenting these draft principles as a *fait accompli*, what follows is meant to be a starting point for further thought and consideration, including philosophical reflection.

[viii] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. https://arxiv.org/abs/1104.3913v2, November 29, 2011. See also S. Barocas and A. D. Selbst. Big data's disparate impact. Technical report, available at SSRN: http://ssrn.com/abstract=2477899, 2014. And more recently, Kroll, Joshua A. and Huey, Joanna and Barocas, Solon and Felten, Edward W. and Reidenberg, Joel R. and Robinson, David G. and Yu, Harlan, Accountable Algorithms (March 2, 2016). University of Pennsylvania Law Review, Vol. 165, 2017 Forthcoming; Fordham Law Legal Studies Research Paper No. 2765268. Available at SSRN: https://ssrn.com/abstract=2765268

[ix] In June 2016, Kabir Alli, an 18-year old graduating senior from a Midlothian, VA. high school, posted a video clip on Twitter of Google image searches for "three black teenagers" and "three white teenagers." The video has been retweeted over 80,000 times to date. Various news media coverage debated the meaning. See http://www.usatoday.com/story/tech/news/2016/06/09/google-image-search-three-black-teenagers-three-white-teenagers/85648838/ and https://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet as samples.

[x] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, PNAS 2014 111 (24) 8788-8790; published ahead of print June 2, 2014, http://www.pnas.org/content/111/24/8788

[xi] See for example: http://www.cbsnews.com/news/researcher-apologizes-for-facebook-study-in-emotional-manipulation/

[xii] See for example: http://finance.yahoo.com/news/social-media-credit-scores-145016363.html

[xiii] The "Trolley problem" has been covered extensively relative to autonomous vehicles; it is difficult to keep up with the increasing media coverage and corresponding fatigue deriving from use of this example. See a recent example here: https://medium.com/enrique-dans/self-driving-vehicles-and-the-trolley-problem-5a3d717b11fa#.122ysq2ia

[xiv] https://en.wikipedia.org/wiki/2010_Flash_Crash