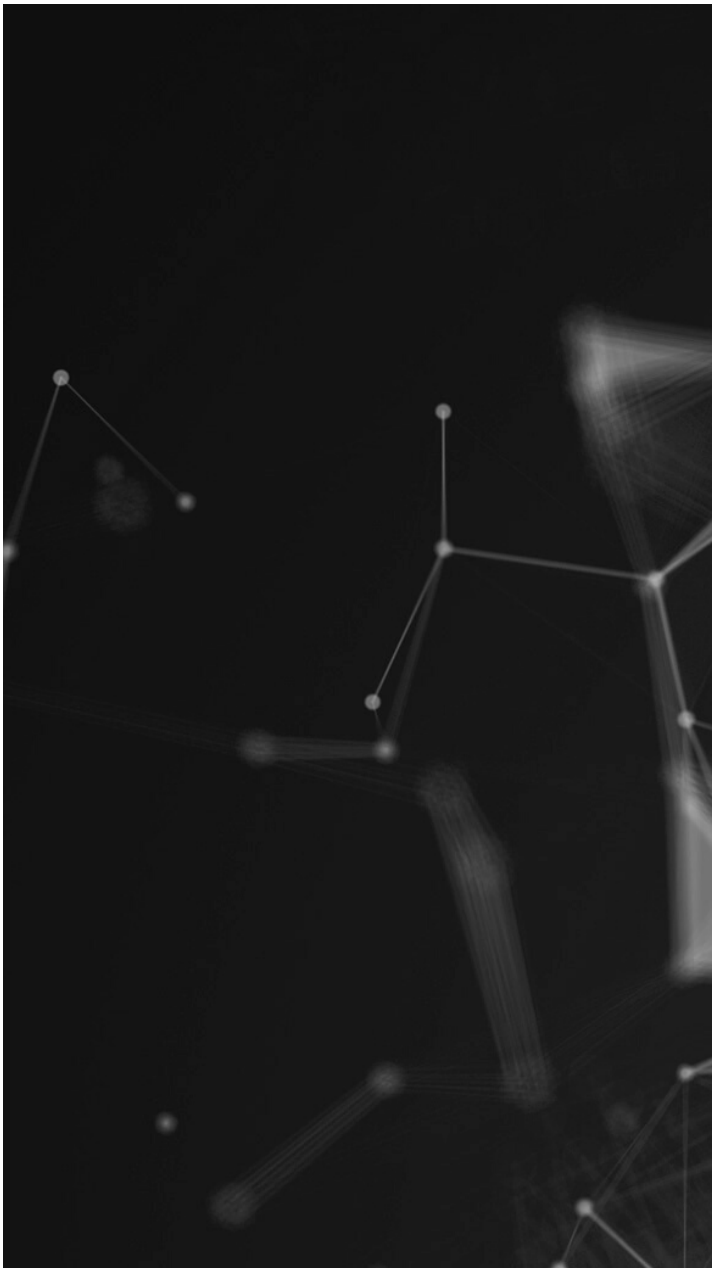
AI ETHICS

A FRAMEWORK FOR THE MEDIA INDUSTRY

Yves Bergquist
yves@etcusc.org
etcenter.org



BACKGROUND

IN THE SUMMER OF 2020, the Entertainment Technology Center (ETC) at USC and the Society of Motion Pictures and Television Engineers (SMPTE) joined forces to create the Joint Task Force for Artificial Intelligence in Media.

THE TASK FORCE'S REMIT is to research opportunities to deploy Artificial Intelligence throughout the media industry, and write a report for SMPTE's standards committee on what standards should be created to accelerate this process.

ETHICS WAS IDENTIFIED EARLY by Task Force members as one of the major areas of focus for its work. This was confirmed by data scientists and media executives during several industry-wide roundtables on AI organized by ETC.

THIS PAPER AIMS TO REFLECT the ideas and suggestions of the group as a whole about what ethical considerations and best practices need to be made in the development and deployment of AI systems in the media industry.

THIS DOCUMENT AIMS TO GUIDE media executives and organizations involved in developing or deploying AI applications. It aims to provide a starting point for an industry-wide conversation about computational ethics as a whole. Many people provided review and feedback to this paper, and we are grateful for their help.

SMPTE AND ETC welcome open debate about one of the most potent technologies of our time. This can be done through ETC's AI Roundtables, which are held remotely every 4 to 6 weeks, and are open to ETC and SMPTE members. Contact yves@etcusc.org to participate. For information about guest participation in SMPTE Standards activities, contact dir.standards@smpte.org.

GET IN TOUCH

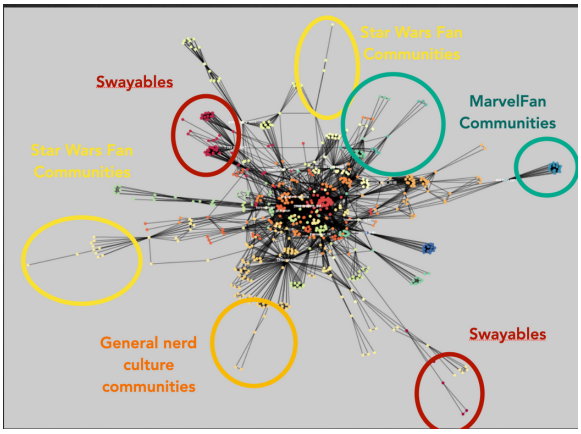
I. INTRODUCTION

(A) WHAT IS AI ETHICS?

AI ethics is the set of ethical considerations involved in the design, development, and deployment of artificial intelligence systems. Because they are hand-built, by humans for humans, AI architectures encode organizational and human biases at all stages of the data science pipeline, from data collection to model deployment.

As such, ethics are concerned with virtually all aspects of AI development: ethical intent in the design phase, respect for privacy and minority representation in data collection, racial, gender, and cultural bias in training data, inclusivity in data science teams, biases in machine models, the need for transparency and explainability, and of course benevolence of end-uses. AI ethics is a complex and evolving ecosystem of practices, systems, and goals. First and foremost, it's an organizational mindset.

In the media industry, this means putting ethics on the agenda of data science and product teams. It means making deliberate and ethically conscious decisions regarding such concerns as data privacy, the use of synthetic media, and content recommendations. It means hiring culturally diverse engineering teams. Communicating transparently with consumers about how their data will be used and decisions about them made. It means training data science teams on identifying cultural biases in data and models, as well as -perhaps most importantly- how to talk about them with business stakeholders.



Influence graph of influential and "swayable" communities of Godzilla fans on Twitter. Source: ETC, 2019.

Direct-to-consumer business models and the considerable opportunity presented by the emergence of Multiverse environments are pushing the industry farther into the arms of machine models, and thus creating even stronger ethical requirements.

As computational intelligence spreads throughout media-making decisions and workflows, it's important for organizations to develop the confidence that their datasets and models are consciously built and fully known, warts and all. This means they are transparent and auditable, trusted, rid of invisible and unwanted biases, and do not result in privacy violations or discriminatory outcomes.

Computer vision models, large language models, and generative models, have all crossed a threshold of performance that carries opportunity and ethical risk. "Deepfakes" arise from a powerful but ethically dangerous technology. Similarly, unbound synthetic characters and agents (chatbots) pose considerable ethical dilemmas.

Ethics is altogether a practice, a mindset, and a conversation. It's an emerging and uncertain, yet essential field. And not just because of ethical risk. Robust, conscious, and transparent data science helps everyone.

(B) WHY SHOULD ETHICS BE USEFUL IN AI DEVELOPMENT?

1. BECAUSE AI IS EARLY, CRITICAL, AND MISUNDERSTOOD. AI is at the same time disruptive, vague, complex, experimental ... and a great story. It's hard to understand and easy to load up with fears and fantasies. This is a dangerous combination.

The convergence of corporate hype, fledgling methods, incomplete and biased datasets, and the urgency to productize, are all a fertile ground for failure. Failure in tech is good, except when when models are put in a position to make decisions about such areas as policing, hiring, synthetic conversations, or even content recommendation and personalization. Then failure comes at a high human cost.

The time to talk about ethical considerations in AI is now, while the field is still nascent, teams are being built, products roadmapped, and minds made up.

Law Firms Turn to AI to Vet Recruits, Despite Bias Concerns

2. BECAUSE IT'S THE LAW. According to the United Nations Conference on Trade and Development, 77% of all UN member states already have data privacy laws or have pending legislation.

GDPR (European Union), and the joint CCPA/CPRA in California have already alerted the private sector on how much attention regulators are paying to consumer data and artificial intelligence-driven decisions. Around the corner, EU's proposed Digital Markets Act makes ethical requirements and data privacy even more ominous.

Closer to us, the City of Los Angeles recently took legal action against IBM for misappropriation of user data for the latter's weather app. Goldman Sachs has been investigated for discrimination against women in some credit card applications. The list goes on, and it will unfortunately get much longer.


nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature › news › article

NEWS | 24 October 2019 | Update 26 October 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

3. BECAUSE FAILING IS EXPENSIVE. As seen above, AI development is no longer just a technical issue, it is increasingly becoming a risk factor. Because AI is altogether experimental, impactful, and expensive, organizations need to examine the downside risk of deploying underperforming and unethical AI systems. Especially because, in most cases, ethical and technical requirements are the same. Unseen bias is as bad for model performance as it is discriminatory, for example. Model transparency isn't just an ethical consideration: it's a trust-building instrument for organizations still viewing AI with suspicion.

In 2017, Amazon had to famously scrap a costly machine-driven job applicant processing piece of software because it discriminated against women (it was trained on an an overwhelmingly male dataset). This cost the company in 3 ways: (1) the obvious reputation hit for a technology leader, (2) the cost of developing, then scraping, a faulty application, but most importantly (3) the opportunity cost of making bad decisions based on biased machine learning models. The next year, an autonomous vehicle tested by Uber killed a pedestrian in Arizona, in part because its model had not been properly trained on jaywalking samples. If we want intelligent machines to make big decisions at scale, we must recognize - and mitigate- the costs of failure.

4. BECAUSE MEDIA NEEDS ITS OWN VOICE. The media and entertainment industry is a tech industry. As such, it has its own voice, its own culture, and nearly 150 years of success marrying human and technological genius. It also holds a substantial and powerful place in our society as the mass distributor of human narratives and social norms.

Media needs to bring this unique voice and hybrid human/machine culture both to AI development and the debate on AI ethics. And as the industry starts developing and deploying AI applications from development to distribution, there is a need to approach this issue at the industry level first.

Media and entertainment companies collect and process large amounts of consumer data, for example. Increasingly this means that they need to comply with a growing list of legal regimes and data governance requirements. Similarly, there's a substantial opportunity to use computer vision in the production (virtual production) and post-production processes (color correction, translation and localization, and of course vfx work).

The quality and diversity of training sets, how color correction can affect representation of minorities, and of course the use of "deepfake" technology, are all critical areas where ethical considerations are paramount. The media industry's history of sophisticated legal practice around likeness rights and participations is a substantial advantage in navigating issues related to computational derivatives of image and content.

At a minimum, the requirements of data and model transparency would go a long way towards reinforcing trust in computational methods and help convert those in the industry still reluctant to use statistical learning to optimize human processes.

Around the corner, the development of conversational agents (chatbots) creates serious ethical risks, especially as the industry looks to create highly immersive and personalized experiences in the multiverse.

Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk

The automated car lacked "the capability to classify an object as a pedestrian unless that object was near a crosswalk," an NTSB report said.

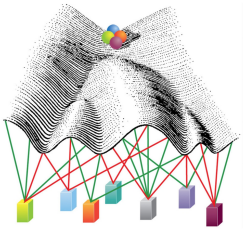
5. BECAUSE IT'S FUNDAMENTAL. As mentioned, technical and ethical standards in AI are overwhelmingly one and the same. Bias is the model-killer. Black box algorithms are inscrutable and can lead to serious unseen bias or underperformance. Intellectual and cultural diversity is critical to high performance in dat science teams. It's healthy for product teams to broaden their system view and consider ethical and societal applications of their work.

Entangling the ethical implications of AI goes a long way towards deepening our collective understanding of the field. And thinking about AI ethics forces us into systems thinking, which is an almost Darwinian imperative in all areas of contemporary business, technology, and society.

(C) SOME CORE PRINCIPLES:

In his excellent book"Trustworthy Machine Learning" (available for free at http://www.trustworthymachinelearning.com), IBM researcher Kush Varshney compares the challenges of trustworthiness in AI and machine learning to those pf processed foods. In the early XXth century, processed food companies like Heinz had to gain the trust of consumers and regulators through "unadulterated ingredients, transparent containers, sanitary food preparation, factory tours, labels, and tamper-resistant packaging". Inspired by this effort to build trust in an essential component of society (food), here are some core principles of AI ethics for the media industry.

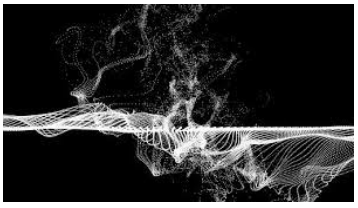
1. BROADNESS



AI ethics need to be viewed in the larger context of computational ethics. Whether or not they are build upon AI or ML architectures (increasingly they are), systems with impact on an organization, a business model, a revenue stream, a key life decisions (such as hiring or getting a mortgage), a minority group, or medical treatment, are subject to bias. As such, they need to be understood, built transparently, and audited regularly. Computational bias, AI or not, means bias on a massive scale.

Ethical considerations should be a systematic part of all aspects of digital product design, development, and QA. This seeding of ethics at the product level is essential to look at bias as a complex ecosystem of inputs, features, models, outputs ... and outcomes.

2. FIT



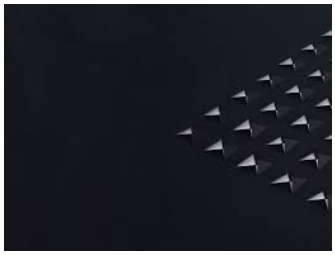
We are what we build. Any organization’s output, products, and decisions (deliberate or not) inherently fits its culture and values. This is why AI ethics is high stakes: it deploys an organization's culture and values on a large scale.

Because they shape society at scale and have a history of taking the public interest seriously, media companies have a distinct responsibility to move forward in their AI ambitions in full awareness of these applications' ethical considerations. They should ensure that all aspects of their development (including data collection), deployment, and end-uses, support the law as well as their own values regarding privacy, justice, tolerance, and human rights.

3. INCLUSIVITY

Gender, racial, social, intellectual, and cultural diversity of all kinds are critical to maintaining a richness of voices, societal experiences, and cultures when developing AI systems. It's not just the right thing to do, it's the smart thing to do. AI and machine learning is extremely hard. It touches upon a large number of technical, academic, cultural and intellectual domains.The more diverse voices, the more viewpoints, the more creative solutions, the more chances of success. Diversity creates richness in products and organizations, and is a critical factor in the performance of data science teams. It's also a good remedy against confirmation bias, which can be costly once enshrined in organizational processes and systems.

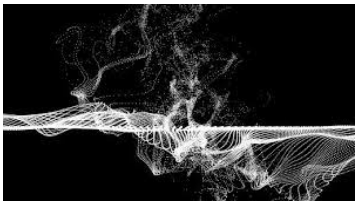
4. TRANSPARENCY AND TRUST



The entire value chain of AI development, from product design to data collection to model deployment, should be secure, transparent, explainable, and auditable. Black box machine learning frameworks are both ethically and statistically dicey. They foster sloppiness in data science teams and mistrust for those already suspicious of machine models. Sure, it's hard to audit the feature representations of each layer of a deep neural net, but that's an audit problem, not a transparency problem. What can't be explained should not be deployed in a decision-making environment.

Only secure, transparent, explainable, and auditable machine models can scale in organizations that are often too suspicious or too enthusiastic. Additionally, all stakeholders deserve transparency, each in their own language, across different points of view and technical sophistication. Ethics should be part of Quality Assurance for any and all computational systems.

5. OPENNESS



AI is still a technical Wild West. Everything around it, from roadmapping to modeling to seeding in company culture, is hard, and will remain so for a long time. Mistakes will happen. Organizations need to communicate comprehensively and with humility about their journey to approach and implement processes around ethical AI. For the benefit of all.

After all, we're all trying to make ethical something that even experts still can't fully understand. Transparency will help regulators, senior business executives, and the general public understand that artificial intelligence is the exact opposite of "magic". It's either blood, sweat, and tears … or it's not AI.

Just like technical and organizational implementation, ethical considerations in the development and deployment of AI are complex and laborious. Being open and didactic will not only feed the public debate about AI with realistic and trustworthy narratives (as opposed to noxious hype), but will create a collective mindshare for organizations to learn from each other's successes and failures.

II. THE AI ETHICS PIPELINE

Implementing the previous principles is where most of the challenge lies. It's harder (and expensive) to audit machine learning models than to build them. Besides, the AI field is still early, and AI ethics is an almost entirely blank slate. Examples of successful, organization-wide implementation of machine learning transparency and trustworthiness are extremely rare. Nonetheless, some early experimentation in the pharmaceutical and financial services industries have suggested some best practices.

(A) ORGANIZATION

This is perhaps the most critical step in laying out an AI ethics strategy, because nothing is more impactful - and difficult to build- than organizational systems, incentives, and mechanisms. This is where AI ethics get enshrined. Here are a few principles borrowed from successful experiments deploying AI governance in a corporate environment:

A. **Set clear goals but flexible roadmaps**. AI ethics is a nascent and uncertain practice that touches upon virtually every business process. It needs flexibility to experiment and diverse buy-in to flourish. It's a good idea to have open and transparent conversations at all levels about expectations prior to setting a roadmap. In media, it means that virtually all sectors across marketing, development, and technical implementation have a piece of the puzzle and a role to play in setting expectations for an AI ethics initiative. Also, the AI ethics work is never done, this will be a perennial trial and error process.

B. **Inventory organizational resources already available to seed an AI ethics program**. Chances are some foundations of an ethics practice already exist within the organization. Set up an executive committee inclusive of all voices, business units, technical backgrounds, and cultures. Promote the initiative (and the group) internally. Educate and train to create a level playing field.

C. **Create clear lines of responsibility and accountability**. Ethics is funded, incentivized and supervised at the corporate level, but its implementation needs to be bespoke to the needs, resources and priorities of each business unit. Product managers in each business unit should be front and center in leading the deployment of ethics policies and practices.

D. **Foster cultural and intellectual variety**. Because of the multifaceted nature of AI ethics, working groups should include a wide variety of stakeholders. This obviously means gender, racial and cultural diversity, but not only. Consumer research teams, product managers, legal and compliance teams, and data scientists all bring a different perspective on balancing the requirements of ethics with that of performance and customer experience. Ethics should not be the exclusive domain of those preoccupied with governance and risk.

E. **Communicate with senior stakeholders**. Learn how to talk about AI and ethics with senior business executives. C Suites and legal executives looking for clear and certain ROI in their AI efforts - including ethics- ignore how experimental the tech still is.

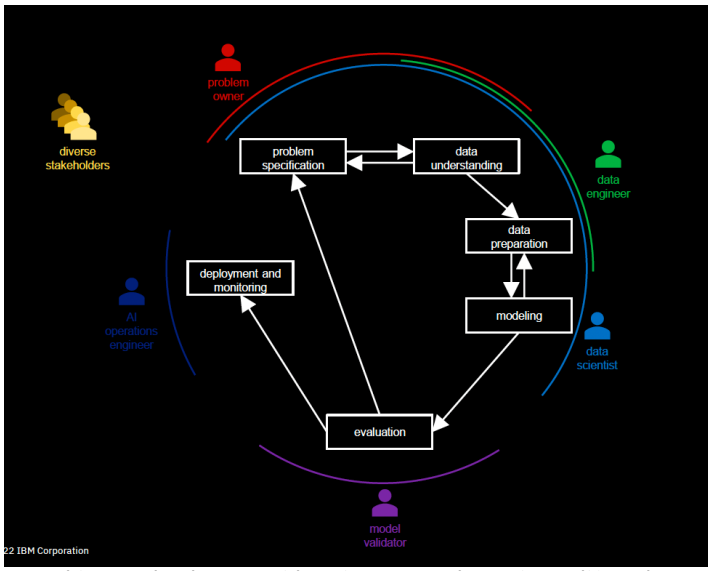
F. **Make the process as transparent and measurable as possible**. Measurement is important in any trial and error process. So is transparency about mistakes and lessons learned with regards to building ethical and trustworthy AI applications. It's a completely new domain related to a completely new technology. Failure will happen.

(B) PRODUCT DESIGN

A. Because computational ethics (not just in AI) need to be implemented as closely as possible to those with responsibility over the use case behind AI applications, two roles within organizations take a pivotal role in AI ethics implementations: insights leaders and product managers.

The first one (internal facing, focused on processes) is straightforward: it is the responsibility of the head of insights to ensure that they are collected, processed, and outputted transparently and ethically. Data and model integrity are a core responsibility.

Product managers could also take a central role in leading external-facing (focused on products) AI ethics considerations. Because they are by nature systems thinkers, care about the customers at least as much (if not more) as about the company, and are ultimately accountable for an organization's *raison d'être* (its products), product leaders are essential "quarterbacks" of computational ethics. They are best positioned to weigh all considerations of transparency, integrity, and functionality. Plainly put, they sit at the intersection of product and users and can best weigh user experience vs ethical requirements.



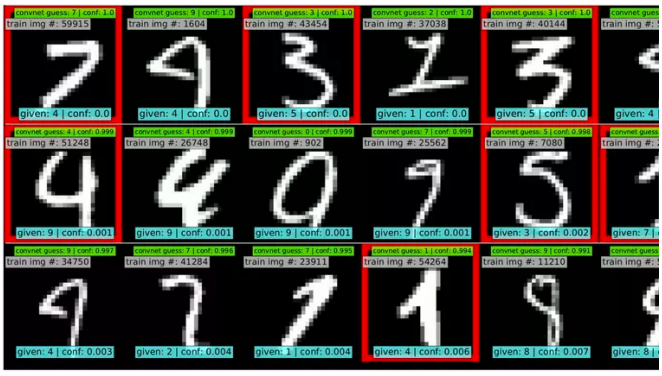
Ethics need to be a consideration across the entire AI lifecycle, according to Dr. Kush Varshney (Dr Varshney's diagram).

Whoever is given the lead to examine ethical considerations in computational systems should start by laying out requirements of "AI trustworthiness and transparency" that are specific to each stakeholder.

A customer will have different needs, and speak a different language, than a marketing executive, or a data scientist. And ethical considerations even vary within customer types. For example, audiences must trust an AI-driven recommendation algorithm to "know" their specific tastes, while intelligently expanding their creative horizon.

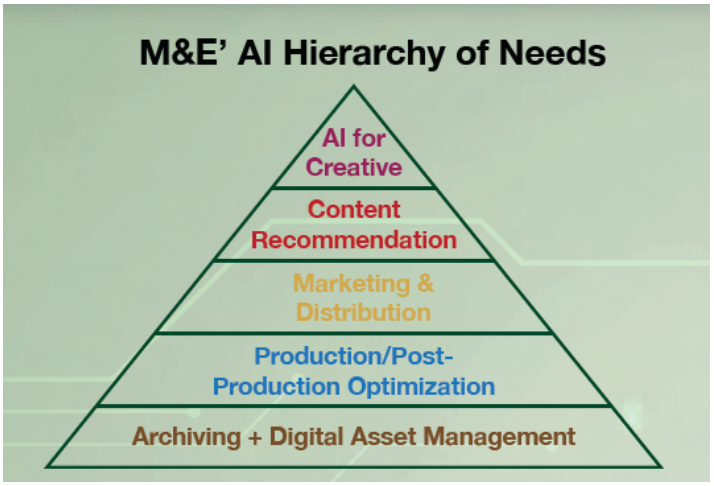
Marketing executives and analysts must trust that a sentiment analysis engine correctly classifies positive from negative sentiment in most cases (deeply semantic sentiment domains like sarcasm are still difficult to measure). A digital product manager must trust that her virtual character won't stray into inappropriate conversations with users.

Listing all stakeholders and analyzing their various cultures and needs is a useful initial step in the AI ethics pipeline. Performance and transparency are also big components of trustworthiness. It's critical to label outputs of AI and ML models with their performance.



Example of a classification model (identifying handwritten numbers) output with confidence levels (source: Kili Technology) .

Identify clearly where technical, business and ethical goals are aligned, and where they are not. Start with he former, and promote quick wins with low hanging fruits (see the hierarchy of needs pyramid). Model performance and sample bias are examples of ethical issues where technical, business, and ethical goals are aligned. Other issues, such as the decision of whether or not to use "deepfake" technology in the VFX process, may require more extensive consultations with C-level executives. Some products may perform better with more intrusive consumer data collection. This could be seen as a user experience problem, where users will have the choice between an enhanced experience that collects data more aggressively and a limited experience that protects privacy.



Finally, an effective AI ethics program will make checking for bias, trustworthiness, and transparency a function of quality assurance. This is a critical part of the ethics pipeline, as it ensures balance between those requirements and the needs of user experience and product performance.

(C) DATA COLLECTION

The data collection process is very much at the heart of the AI practice as a whole, and of ethical considerations in particular. "Garbage in, garbage out" is the Golden Rule of data science: models are only as good as the data they are trained on. Identifying biases in data collection (and monitoring how bias might increase over time) is at the heart of the AI ethics practice.

1. Know your data

● This is perhaps the most important part of the AI ethics pipeline. It's also a major area where statistical and ethical requirements are one and the same.

● Data is the raw material of data science, and it is data scientists' first and foremost responsibility to know their dataset, its strengths and weaknesses, inside and out, to be able to map issues with a skewed output (the model) back to skewed inputs (the data). Use representative datasets that fully take into accounts gender, race, culture, etc. This is not just the right thing to do, it's the statistically sound thing to do.

● Sample bias is a primary source of poor ethical outcomes in AI. For example, facial recognition applications have been notoriously underperforming in the detection of both darker skin tones and females (see Dr. Joy Buolamwini's and Dr. Timnit Gebru's "Gender Shades" study), due to substantial under-representation of darker skinned samples in computer vision training sets.

● In their 2018 paper, Gebru and Buolamwini noticed that two of the most prominent training sets of faces at the time, IJB-A and Adience, are composed of 79.6% (IJB-A) and 86.2% (Adience) of lighter skinned faces. They found that the maximum error rates for lighter skinned males in these models was 0.8%, vs 34.7% for darker-skinned females (http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf). A subsequent 2018 study (https://arxiv.org/abs/1812.00099) dove further into other standard datasets and found a different explanation, which illustrates the need to be intimately familiar with one's data.

- Similarly, the bias in Amazon's hiring software came from the fact that it had been trained on resumes received by the company over a 10-year period. An overwhelming majority of these resumes were from men. As a result, it penalized resumes that included the word "women's", as in "women's basketball team".

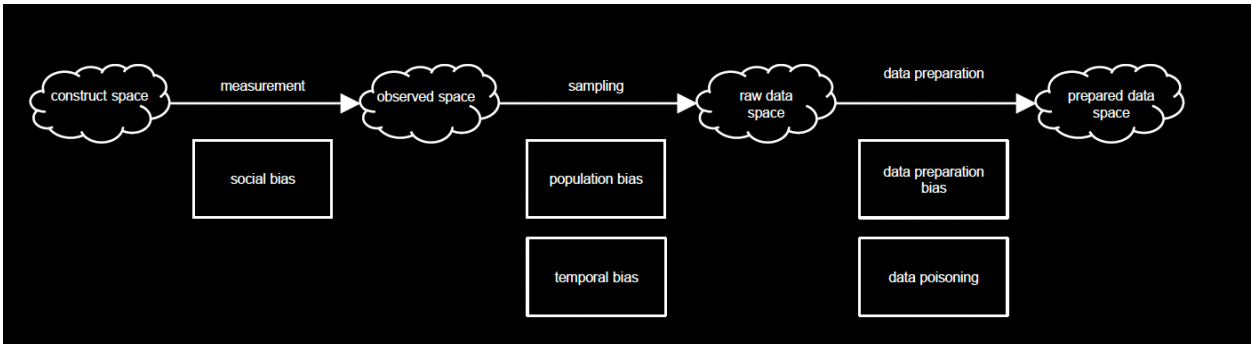
- This is a complex and very multi-layered issue to tackle. For example, even very subtle nuances in how data is collected (the way a question is asked, or how incentives to contribute one's data are structured) that can dramatically affect the resulting dataset. Knowing how the data has been collected, and what biases may lie within that process, is increasingly a core responsibility of data scientists.

2. Know your problem

- AI and machine learning are used to solve real world problems by using data to represent and model those problems in their larger context. This is identical to how the human brain functions: we use data to model (summarize) an infinitely complex world, and use those models to act upon the world. Knowing intimately the problems, systems, behavior, phenomena they are trying to model, and in this case, what biases may be inherent in them, is a key strength of great data scientists. Some parts of our world are simple, but most are extremely complex (not least of which human behavior). When modeling a real-world system (such as audience decisions) data scientists need to understand is the set of variables they are analyzing accurately represents (e.g., sums up) the larger system they are trying to generalize their findings about. They also need to identify if and when the inequities or biases of the system itself will get passed on to the model. Often times the set of variables available to the system is too small and partial for the model to generalize to a much more complex (and fast evolving) real-world system. Over and under-fitting are the statistical manifestations of this issue, as is encoding real-world biases in machine models.

- This is especially important when modeling human behavior, specifically during audience research. Posting a thread on Reddit or retweeting a tweet on Twitter or liking a post on Facebook are 3 radically different kinds of social behavior expressed by different genders, age groups, races, and sub-cultures. And models about them generalize differently, which is why it's critical, in the practice of AI ethics, to maintain intimate knowledge of how underlying social, cultural or behavioral biases may impact data collection. It's critical for data scientists to understand the underlying biases not only in the input data but in the systems they are trying to model.

- This is best done as a collective process, since confirmation biases (the tendency to look for, cherry pick, or interpret insights according to one's preexisting beliefs) are also present in data science teams, or quantitatively-driven functions such as marketing and consumer research.



Dr. Kush Varshney 's representation of the many steps in the AI lifecycle where bias can be introduced.

3. Communicate clearly about use (opt-in), biases, and their tradeoffs

- When collecting user data, opt-in is a must (it's also increasingly the law). The best practitioners in this domain avoid legal language and use instead a simple user-facing explanation of how personal data will be used, and what the implications of opting in and out are for the user experience (for example, an opt-out of sharing data would affect personalization).

- Bias often can't be avoided, in which case it's critical for data science teams to communicate fully and clearly to end-users about their model's skews resulting from the bias. A simple annotation in the output can be very powerful in building transparency and trust, without which no culture of data (let alone AI) can be successfully scaled in media organizations.

(D) MODELING

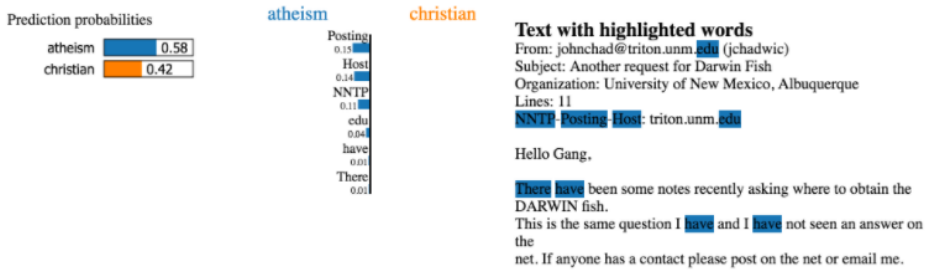
Traditional measures of performance In ethically compliant AI or machine learning, building trust is both critical and labor-intensive. There are two parts of this: creating explainable statistical models (model transparency) and effectively reporting key features of the model, so it can be quickly audited by end-users for bias and potential performance variations.

1. Model Transparency

Transparency means building simple models to explain complex ones, to give data scientists a window into often complex and inter-locking machine learning architectures. This is increasingly a challenge, as neural net architectures become deeper and more integrated with other types of models. Luckily, the past few years have seen a flurry of development of AI transparency tools. All providers of cloud-based machine learning have started offering tools to interpret and understand their models (for example, AWS's Sagemaker model explainability feature - based on SHAP-, Microsoft's InterpretML toolkit, or Google AutoML's feature scoring tool). These are useful because variables in a mode are hierarchical (some are more powerful than others within the model), and surfacing that structure is a key step in understanding how the power of certain variables related to gender, race, or culture, for example, may perpetuate inequalities.

Here are some good and popular explainable AI tools:

- SHAP (SHapley Additive ExPlanations): this is perhaps one of the most widely known AI transparency tools, because it works across a wide range of models, from linear regression to deep learning, and covers a wide variety of domains including computer vision and NLP. SHAP uses a game theoretic approach to rank features (for example, words in a sentiment analysis model) by order of importance in predicting the output. This is the kind of hierarchical view that is very helpful not just in the context of AI ethics, but for data scientists to QA their own models. Transparency one of many areas serving both ethics and model performance.



LIME model on a text string

- LIME (Local Interpretable Model-Agnostic Explanations): similar to SHAP, but more computationally effective (quicker). LIME also ranks features by ow much it contributes to the output (in an image classifier it can produce a heat map of an image with “useful” features in green and “not useful” features in red - SHAP can do this as well). This is very popular for Python’s scikit-learn users because of its built-in integrations.

- ELI5: this works similarly to SHAP and LIME but is perhaps the most popular transparency packages in Python, because of its integrations across the board with scikit-learn, XGBoost, Keras, and a few more.

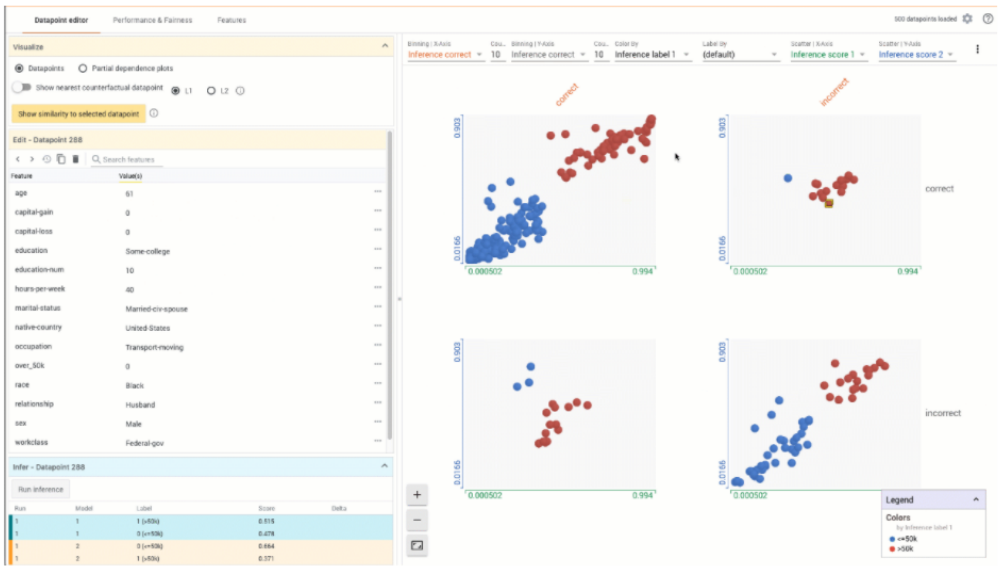
Out[10]: **y** (score **3.748**) top features

Contribution[?]	Feature	Value
+1.340	mean concave points	0.038
+0.883	<BIAS>	1.000
+0.739	worst concave points	0.102
+0.521	worst area	677.900
+0.510	worst texture	24.640
+0.389	compactness error	0.019
+0.306	worst perimeter	96.050
+0.181	area error	30.290
+0.171	mean texture	18.600
+0.066	symmetry error	0.018
+0.047	perimeter error	2.497
+0.026	concave points error	0.010
+0.026	fractal dimension error	0.004
+0.025	mean radius	12.470
+0.023	mean compactness	0.106
+0.016	mean fractal dimension	0.064
+0.010	radius error	0.396
-0.072	smoothness error	0.007
-0.136	worst radius	14.970
-0.167	worst symmetry	0.301
-0.187	mean smoothness	0.100
-0.279	worst smoothness	0.143
-0.690	worst concavity	0.267

Example output from ELI5

- AIX360: developed by IBM Research but still open source, this toolkit has extensive functionality and isn’t dissimilar from Google’s “what if” tool, but can be used outside of the Google CloudAI environment, although it is not for beginners.

- Google’s “What-if Tool”: this allows data scientists to test a model's performance under a variety of different situations. It helps understand the impact of various variables (such as race or gender) on the model itself. This is an excellent and intuitive tool for beginners using the Google Cloud AI infrastructure, as it has a very good visual interface and can be run easily (and with minimal code) from platforms such as Jupyter Notebooks, Google Colab, and even Tensor Flow's TensorBoard dashboard. It can be used at various stages of the data science workflow. It can support TensorFlow models out of the box. Works with tabular, image, and text data.



Screenshot of Google's "What if Tool" dashboard

Transparency is not just key: it's a perennial concern. The world changes, the problem changes, the data changes, and model performance is affected. There is no longer a fit between the model and the system, or behavior, it's representing. "Model drift", as it's called in data science circles, impacts ethical outcomes, because what may be ethical in January may no longer be in June. Only transparent and auditable models can catch model drift before it causes damage.

2. Model reporting: model cards

Too often, machine models are released with incomplete documentation. As a result, they are applied to contexts in which they do not perform well or are not appropriate to deploy in. Or they exert substantial yet hidden influence on decisions that affect us as digital citizens. This is a rising concern in public policy, where local and national governments are starting to expose computational decisions and how they are made. Recently, the cities of Helsinki of Amsterdam have created a public online registry laying out in detail the algorithms, models, and data used to make specific public decisions. It's only a matter of time until this becomes a widespread rule in the public sector.

There's an interesting framework for this. Created by Dr. Margaret Mitchell and Dr. Timnit Gebru's team at Google in 2018, "model cards" are standardized documentation laying out all the information necessary to evaluate a model and benchmark its performance in a variety of contexts.

Sure, libraries and models often come with documentation, but it's often incomplete, too long, and generally could use standardization. Model cards are standardized "food labels" for data science that also -ideally- benchmark a model's performance in a variety of contexts and use cases, some related to inclusion and bias. Per the Mitchell/Gebru/team paper ("Model Cards for Reporting": https://arxiv.org/pdf/1810.03993.pdf):

> "Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information".

Data Label

Data Transparency Facts

Data Distributor Name: **Data Company**
Data Distributor Contact: DataSolutionTeam@data.com
Data Provider Name: **Leasing Company**
Data Provider Contact: DataAccounts@leasingco.com

Audience Snapshot	
Branded Name	Auto Intenders – Six Months
Standard Name	Auto Intenders
Audience Description	Households likely in the market to purchase a new vehicle in the next six months
Geographies	USA

Audience Construction	Attributes
Audience Count	6,500,000
Precision Level	Households
Activation ID(s)	Cookies
Audience Expansion	Yes
Cross-Device Expansion	Yes
Last Refresh Date	02-Jan-2018
Event Lookback Window	60 Days

Data Source	Attributes
Source ID Description	Dealer-reported names and postal codes of individuals who requested test drives
Source ID Contribution	1,130,000
Precision Level	Individual
ID Key	Name and Postal
Source Event	Transactions
Inclusion Method	Observed
Seed Size (if modeled)	-
Source Refresh Frequency	Quarterly
Event Lookback Window	180 Days

This Data Transparency Label has been developed by members of ANA's Council for Data Integrity and IAB Tech Lab's Data Transparency Working Group, with the support of CIMM, The ARF and IAB's Data Center of Excellence. For more information, please visit datalabel.org.

Model card example

(https://iabtechlab.com/press-releases/major-advertising-trade-bodies-unveil-data-transparency-label/data-label/)

Model Card for Census Income Classifier

Model Details

Overview

This is a wide and deep Keras model which aims to classify whether or not an individual has an income of over $50,000 based on various demographic features. The model is trained on the UCI Census Income Dataset. This is not a production model, and this dataset has traditionally only been used for research purposes. In this Model Card, you can review quantitative components of the model's performance and data, as well as information about the model's intended uses, limitations, and ethical considerations.

Version

name: 36dea2e860670aa74691b5695587afe7

Owners

- Model Cards Team, model-cards@google.com

References

- interactive-2020-07-28T20_17_47.911887

Considerations

Use Cases

- This dataset that this model was trained on was originally created to support the machine learning community in conducting empirical analysis of ML algorithms. The Adult Data Set can be used in fairness-related studies that compare inequalities across sex and race, based on people's annual incomes.

Limitations

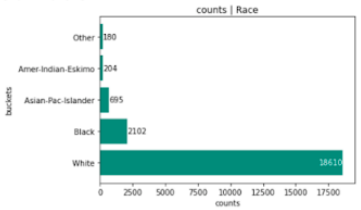
- This is a class-imbalanced dataset across a variety of sensitive classes. The ratio of male-to-female examples is about 2:1 and there are far more examples with the "white" attribute than every other race combined. Furthermore, the ratio of $50,000 or less earners to $50,000 or more earners is just over 3:1. Due to the imbalance across income levels, we can see that our true negative rate seems quite high, while our true positive rate seems quite low. This is true to an even greater degree when we only look at the "female" sub-group, because there are even fewer female examples in the $50,000+ earner group, causing our model to overfit these examples. To avoid this, we can try various remediation strategies in future iterations (e.g. undersampling, hyperparameter tuning, etc), but we may not be able to fix all of the fairness issues.

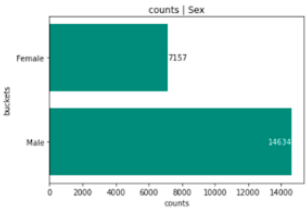
Ethical Considerations

- Risk: We risk expressing the viewpoint that the attributes in this dataset are the only ones that are predictive of someone's income, even though we know this is not the case.
  Mitigation Strategy: As mentioned, some interventions may need to be performed to address the class imbalances in the dataset.

Train Set

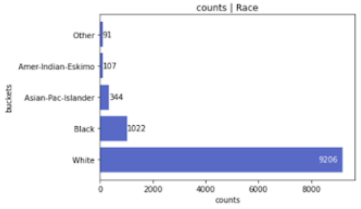
This section includes graphs displaying the class distribution for the "Race" and "Sex" attributes in our training dataset. We chose to show these graphs in particular because we felt it was important that users see the class imbalance.

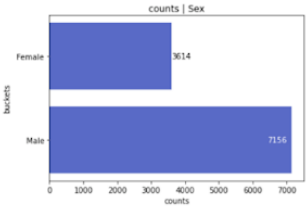




Eval Set

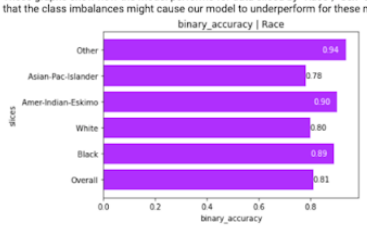
Like the training set, we provide graphs showing the class distribution of the data we used to evaluate our model's performance.

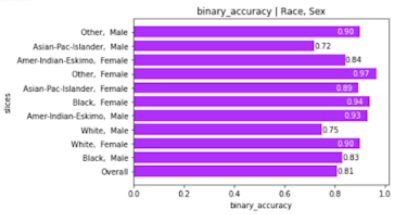


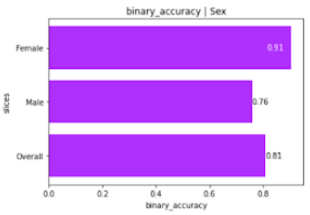


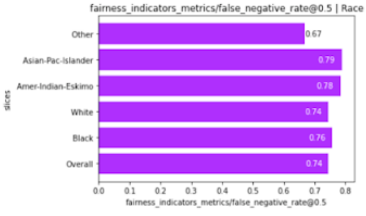
Quantitative Analysis

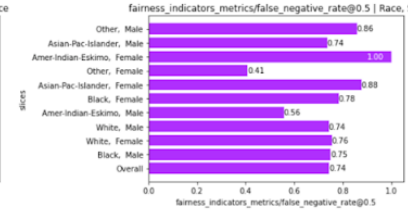
These graphs show how the model performs for data sliced by "Race", "Sex" and the intersection of these attributes. The metrics we chose to display are "Accuracy", "False Positive Rate", and "False Negative Rate", because we anticipated that the class imbalances might cause our model to underperform for these metrics.

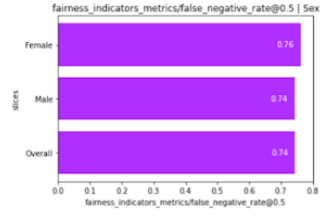


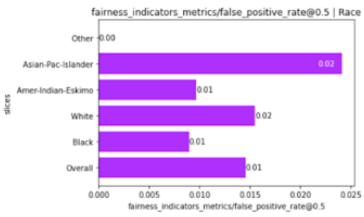


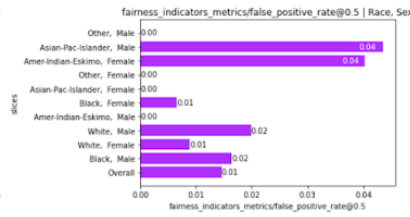


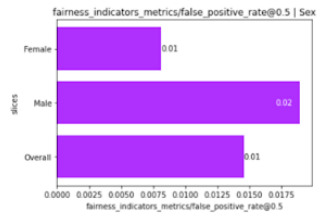




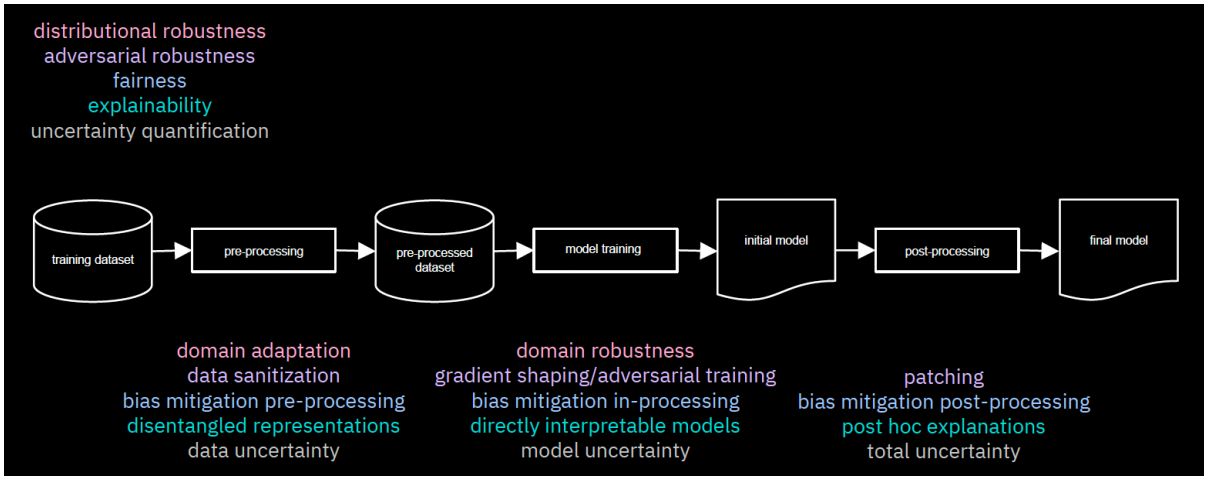








Another Model Card example (Google)



Dr. Kush Varshney's representation of the ethical and trustworthiness needs during modeling

As stated consistently in these pages, there's no real playbook for AI ethics. The technology is early, and its ethical considerations another area where the hype has vastly preceded the reality. Even regulators have, by and large, limited their interest to data privacy - for now.

But because the technology is so complex and increasingly important, and its presence is so ubiquitous, it needs to be approached and roadmapped through multifaceted systems thinking. There are simply too many components in any serious artificial intelligence effort to avoid considering its requirements and ramifications - especially ethics- as anything but a system.

The most successful organizations in building and scaling AI internally are the ones that think about it the most thoroughly and systematically. And nothing forces an organization to think deeply - and in systems- more than ethics. Far from window-dressing or virtue signaling, putting ethics front and center will bring about the modes of operation, intellectual rigor, and organizational culture necessary to excel in building AI systems.