GRADIENT INSTITUTE

https://gradientinstitute.org
info@gradientinstitute.org

Ethical AI Education

Version date: 21/02/2020

Executive Summary

Organisations have recognised the challenge of creating ethical AI systems, and are in the process of developing frameworks for data ethics and the ethical use of AI. However, a gap remains between defining high-level principles such as 'minimise harm' or 'be fair', and building an AI system that embodies them.

Gradient Institute is a not-for-profit organisation of AI experts devoted to closing that gap; to bringing ethical AI systems into practical use as widely as possible. Doing so is a complex and unsolved problem, but progress is being made by a community of researchers and practitioners around the world.

Gradient Institute develops and runs education courses designed to show organisations the forefront of this progress. Gradient Institute's courses are designed for data scientists, engineers, domain experts and decision-makers, and provide the technical and non-technical concepts and skills needed to begin making real AI decision systems more ethical.

1. Motivation

AI systems increasingly provide automated decision making or decision-support that potentially impacts the lives of millions of citizens. Such systems can be more accurate, consistent, scalable, and auditable than human decision-making alone.[1]

However, the nature of AI means that its use of context, and the way it makes decisions is fundamentally different from humans. To ensure that AI-driven decision systems behave ethically in accordance with their designer's intent, these differences must be accounted for. Doing so is a new and difficult challenge, with open questions not yet answered even theoretically.

Organisations have recognised this challenge of creating AI systems that behave ethically, and some are in the process of developing, or have already completed, frameworks for data ethics and the ethical use of AI. However, a gap remains between high-level principles such as 'be fair', and actual AI systems that embody that goal.

In reality, even a statement like 'be fair' is likely to be impossible to satisfy, for a simple reason: there are many different ways any decision process can reasonably be said to act fairly, and in the vast majority of cases, these different kinds of fairness cannot all be satisfied at the same time. Note too that this impossibility holds irrespective of bias in the data or whether a human or machine is involved.[2]

A more realistic goal is the following: remove any particular forms of discrimination forbidden by law, remove *needless* discrimination caused by biased data or mistakes in design, then understand the different kinds of remaining differential treatments and the different groups they apply to, and finally, thoughtfully *trade off* levels of differential treatment and outcomes with the primary goals of the system. This is the difference between the high-level principle 'be fair' and reality. It is a complex challenge, and group fairness is only one example of the many difficult practical and theoretical issues that a designer must solve to bring their ethical intent into reality.

Moving from a set of ethical principles to a working system requires the combined effort of data scientists, their leaders, and the senior decision-makers in the organisation. It requires eliciting, in precise detail, legal constraints, the trade-offs between different goals of the system, and different ethical concerns. It requires understanding the cause and effect relationships between the system's

[1] See for instance, Kleinberg, J., Ludwig, J., Mullainathan, S., Sunstein, C. *Discrimination in the age of algorithms*. https://arxiv.org/abs/1902.03731

[2] See, for instance, Kleinberg, J., Mullainathan, S., Raghavan, M., *Inherent trade-offs in the fair determination of risk scores*. https://arxiv.org/abs/1609.05807

actions and their impacts in the world, in both the short and long term. And it requires implementing a regime of constant and ongoing monitoring and validation of these systems to ensure they continue to perform as intended.

Gradient Institute believes it is these skills organisations using AI must next acquire. We have designed training courses to begin that process: technical courses for data scientists and data engineers with hands on the data and tools, and conceptual courses for leaders to understand the issues and their responsibilities in defining the proper intent and trade-offs of the system.

2. Case study: Targeted Selection

Consider the problem of selecting individuals to assess via an audit, for example, a government performing trying to verify compliance with relevant tax law. Modern implementations of such a program typically estimate the risk of non-compliance with a machine learning system that is trained on historical audits. Individuals that the algorithm predicts are 'high risk' are selected for auditing.

Equivalent selection problems are common in industry, for example, selecting customers likely to move to competitors in the future for special discounts or loyalty problems, or to determine whether a customer should receive a home loan or credit-card based on their predicted likelihood to repay. The issues raised in the auditing example equally apply in these cases.

The targeted selection problem is a simple example of an AI-driven decision system, but it is still replete with ethical challenges. These challenges touch every aspect of such a system, from data collection, to modelling and prediction, to decision-making, and the feedback loops created by the system's actions. To properly address these ethical issues requires input from many parts of an organisation, not just data scientists, but also their leaders and senior executives.

The following list enumerates just some of the specific issues that would need to be addressed in a targeted selection problem, in this case framed in the government tax auditing scenario described above.

Data Gathering

- The historical data being used to model risk is unlikely to be a random sample of the cohort in question. Far more likely is that an older targeting regime was used to assess risk and select based on that risk. Failing to properly account for this with a suitable model can lead to biased predictions, especially if there was any sort of intentional or unintentional bias in the historical targeting.

- Using historical data in a predictive model implies that the underlying relationship between the available data and features does not change over time. This is unlikely to be true in real problems and must also be accounted for.
- Real data about real people often has gaps: missing data that was either not provided or not collected. Unfortunately, missing data are rarely distributed randomly. In many real-world examples it tends to correlate (positively or negatively) with socio-economic disadvantage. This means that imputing that missing data for the purposes of modelling can introduce additional bias to the final decisions.
- The disadvantaged tend to be either over- or under-represented in demographic datasets. Any data related to crime, welfare, business, or geographical remoteness for example would likely show this effect. Yet more care is therefore needed to account for this effect and not bake-in discrimination.

Risk Estimation

- All else being equal, models are rarely equally accurate across all divisions of society. For example, they may be more accurate on men than women even though they are represented equally in the dataset. As a result, even data that faithfully reflects reality (i.e., devoid of sample selection bias) leads to predictions that will tend to reinforce existing inequities in general unless active measures are taken by the designers.
- Many of these systems apply targeting decisions in an on-going or iterative way. This means that they are interacting with the people on which they're making predictions, and therefore changing the system. In other words, decisions made to target an individual have a causal effect that means historical data no longer represents the true cohort. Accounting for this causal feedback loop is vital in producing an AI system that acts as intended.

Selection

- Even with an accurate estimation of an individual's 'risk' score, the decision process of selecting which people to audit is complex. Many considerations must be taken into account such as:
    - *False-positive / false-negative cos*t: The impact of a wrong prediction is complex in real situations. Putting someone through an unnecessary audit or examination is costly to them, but failing to audit individuals that are non-compliant costs government and often society as well.

    - *Law*: How does anti-discrimination and other applicable law limit the set of possible selections and selection strategies? Certain types of differential treatment may, for example, be at risk of unlawfully discriminating.
    - *Fairness*: How are the risk scores distributed across different demographics, and how are the mistakes made in predicting risk scores distributed? Having the majority of 'false positives' come from a particular ethnic group for example systematically discriminates and likely has long term negative consequences.
    - *Exploration / exploitation trade-off*: As well as finding individuals, the targeting is also choosing the new data points that will subsequently be used for training. This must be explicitly accounted for, otherwise the resulting feedback loop will increase bias in the system.
- All these factors: the financial cost and benefit of the scheme to the government, the long-term cost or benefit to society, the differential impacts across different demographic groups and the individual harms and benefits to those targeted, cannot be simultaneously optimised. They exist as inescapable trade-offs. These trade-offs need to be understood precisely, and then often difficult calls must be made by senior leadership to determine what balance is appropriate.
- These trade-offs are not static: as the system evolves, so too will the balance of compromise required. This means the system must be constantly monitored and adjusted as necessary.
- The trade-offs are also probabilistic, requiring an understanding and commitment to a certain level of risk.

The take-away message from this case-study is that there is a vast chasm between high-level ethical principles for AI systems and those systems working in production. It is also clear that an 'ethics panel' or similar conceptual analysis of the system at the early design stage will never be enough to ensure the system acts ethically. What will determine how ethically a system behaves is a complex function of its detailed design and implementation, the particular statistical properties of the data and models used, and how it is monitored and adjusted over time.

3. Gradient Institute Courses

Gradient Institute has developed four courses in ethical AI targeting different levels of an organisation:

1. A two-day **data scientist's course**, aiming to develop the technical skills necessary for building systems that use machine learning to make automated decisions whilst accounting for ethical objectives.

2. A one-day **practitioner's course**, aiming to provide an understanding of some of the theoretical, technical and organisational challenges in creating ethical AI systems, the roles and responsibilities of data scientists, domain experts and leaders, and what information needs to

come from, and go to, senior strategic decision-makers. It covers similar content to the data scientist's course, but at a more conceptual level. This course will be offered from Jan 2020.

3. A three hour **leader's course**, aiming to introduce the key challenges in building ethical AI systems from the perspective of leadership. The course highlights responsibilities that leaders have for determining trade-offs between ethical and other objectives, and for creating a culture of rigour and accountability around the design, use and monitoring of AI decision-making systems.

4. A two hour **board/executive course,** aiming to introduce the key risks (including ethical considerations) in building, procuring and operating AI systems and provides strategies for managing these risks. The course has been designed for boards and senior executive teams and can be customised for the organisation. If two hours is not available, a one hour version can be delivered (though we recommend the two-hour version as there is a lot of material of relevance to boards and executives).

Please note that these courses are designed to be mutually exclusive: Gradient Institute advises against having a participant attend more than one of them. The following sections describe these three courses in more detail.

3.1 Data Scientist's Introduction to Ethical AI (coding)

Intended audience: People that can (or do) build data-driven model pipelines professionally.

Duration: Two days.

Prerequisites: Expertise in a data-driven modelling discipline, either machine learning or statistics. Proficiency in using a vector based programming language (such as Python & numpy or R) to implement predictive modelling pipelines.
Note: Gradient Institute administers a short quiz to determine if this course will suit students, as it involves substantial hands on coding activities. The practitioners course provides an alternative, higher level overview of the same concepts without the coding requirement.

Format: Participants work through Jupyter notebooks with exercises and examples under the tutelage of Gradient Institute data scientists. The tutors will provide a short presentation at the beginning of each section and will lead class discussions. Gradient Institute provides the notebooks to participants for their reference after the course. This course is run in groups of up to fifteen students with two Gradient Institute tutors.

Outcomes: By the end of the course, participants will have built simple example systems that use machine learning to make automated decisions whilst accounting for ethical objectives. Participants will understand and explore some of the technical pitfalls that prevent machine learning systems from behaving ethically, and how to identify and correct for them.

Outline of topics:

- *Robust Modelling & Dataset Shift:* Machine learning (ML) models are built on strong assumptions about their data, that real systems may violate. Failure to recognise these risks can lead to unintended and unethical outcomes. We examine the limits of ML and cross validation, and when pathological circumstances can be addressed with alternative approaches.
    - What is dataset shift? What harm can it cause?
    - Detecting dataset shift (covariate shift) and what can be done about it.
    - What outliers in the data are, and what can they do to a machine learning model.

- *Causal versus Predictive Models:* ML models leverage correlations in data to predict outcomes, on the assumption that the data generating process is fixed. Where models are used to drive decisions and interventions, failure to consider causality can lead to poor consequences despite good intentions. We clarify the distinction between causal and predictive models and how they can be used & interpreted.
    - Identifying when a causal model is required
    - Understanding Simpson's Paradox
    - Feature importance, partial dependence and causality

- *Loss Functions:* Building a data-driven automated decision system requires explicitly specifying its objectives, often in the form of a loss function and constraints. The particular choice of loss, including what considerations to omit and include, are the primary mechanism of control designers have over the ethical operation of the system.
    - Expected utility
    - Cost-sensitive classification
    - Quantifying inherent trade-offs in prediction problems

- *Fair Machine Learning:* ML systems can perform well on average and still systematically err or discriminate against individuals or groups in the wider population. We examine some of the common notions of algorithmic fairness that attempt to measure and correct for such disparate treatment or outcome in ML systems.
    - Sources of unfairness in machine learning models
    - Fairness metrics
    - Approaches to removing bias

- *Interpretability, Transparency & Accountability:* These approaches help us identify when models might break down, when they are lacking vital context, and whether they have been designed and motivated in an acceptable way. In combination, they can make can help us be more accountable for the decisions our systems makes. We provide an introduction to some of the tools and techniques available for making models more interpretable and transparent.
    - Motivations & audience for interpretability
    - Feature importance and partial dependence
    - Local interpretability & LIME
    - Global interpretability & surrogate models

- *Project:* The final component of the course is an applied project that challenges students to put the concepts learned into practice. Students will work in teams to tackle a short data science project and present the results to the group at the end of the day.

Gradient also provides some refresher material available online for participants to study before the course:

- *Python for data science:* We provide a brief review of the key tools & techniques for data science, primarily for students who primarily use another modelling language such as R, SAS or STATA.
    - Interactive analysis in Jupyter notebook
    - Numpy: working with arrays
    - Pandas: working with tables
    - Sklearn: a framework for machine learning

- *Supervised learning & model validation:* We provide an overview of the theoretical foundations of machine learning and model validation, primarily as revision but with an emphasis on ensuring a strong conceptual understanding.
    - Core concepts underlying supervised learning
    - Overfitting & underfitting
    - Model uncertainty

- Classification & regression
- Predicting and calibrating probabilities

3.4 Practitioner's Introduction to Ethical AI (no coding)

Intended Audience: People who build models, but are not confident with the coding required for the Data Scientist's course. People who use or interact closely with models and the technical teams building them. This includes domain experts and policy staff using or planning to use models as part of their work, as well as managers of technical teams and people responsible for oversight of AI systems.

Duration: One day.

Format: Interactive workshop. A combination of working through interactive Jupyter notebooks, presentation and discussion. This course is run in groups of up to fifteen students with two Gradient Institute tutors.

Prerequisites: Experience in quantitative reasoning. Exposure to machine learning or statistical modelling.

Outcomes: Participants will gain a conceptual understanding of some of the theoretical, technical and organisational challenges in creating ethical AI systems, the roles and responsibilities of data scientists, domain experts and leaders, and what information needs to come from, and go to, senior strategic decision-makers.

Outline of Topics: The same topics as the data scientist's course above with the addition of an introductory module on automated decision systems. Topics are explored at a more conceptual level, without requiring participants to write code.

3.2 Leader's Introduction to Ethical AI

Intended Audience: Leaders and strategic decision-makers.

Duration: Three hours.

Prerequisites: None.

Format: Interactive presentation from Gradient Institute staff interspersed with discussion and activities.

Outcomes: Participants will understand the key challenges in building ethical AI systems from the perspective of leadership. They will see the responsibilities that leaders have for determining trade-offs between ethical and other objectives, and for creating a culture of rigour and accountability around the design, use and monitoring of AI decision-making systems.

Outline of Topics: After an introduction to the concepts of ethical AI, participants will explore some of the key considerations for organisations that are designing, implementing, maintaining and governing ethical AI systems. These include

- *Quantifying Intent* - AI systems require precise, mathematical objectives specified in terms of measurable quantities. We examine the challenge of defining these objectives, and some strategies to help ensure that the operation of the resulting system is aligned with its designers' intent.
- *Modelling Impact* - Typical AI and machine learning systems are concerned with discovering useful correlations for prediction in previously acquired training data. However, this training data may not account for the consequences of the AI system interacting with the world, a key question when understanding the system's ethical impact. We discuss the *causal* approach to modelling needed in this situation, and how designers might recognise when it is necessary and when a simpler correlational approach is sufficient.
- *Balancing Objectives* - It is unlikely that an AI system will be able to satisfy all of its objectives simultaneously. We explore the process that those people responsible for the system must undertake of making fundamentally subjective but consequential choices about the degree to which each objective is satisfied. We also discuss the unavoidable set of trade-offs between different notions of fairness, and of the system's accuracy.
- *Testing and iterating* - Building a trustworthy AI system requires careful testing and iteration of the mathematical, computational and organisational aspects of the design and implementation. We explore this assurance challenge and some approaches with which to address it.

- *Responsibility and Governance* - Systems that make decisions automatically must still have a human accountable for their actions, and for the design decisions that govern those actions. We discuss empowering the right people to be responsible for an AI system, and the need to ensure they have the tools and ability for that responsibility to be meaningful.

3.3 Board/Executive Introduction to Ethical AI

Intended Audience: Boards and senior executive teams.

Duration: Preferably two hours (but a one-hour version can be delivered if necessary).

Prerequisites: None.

Format: Interactive presentation from Gradient Institute staff interspersed with discussion.

Outcomes: Participants will understand the key risks (including ethical risks) in building, procuring and operating AI systems and have strategies for managing these risks. They will also see the responsibilities that the Board and executive have for determining trade-offs between ethical and other objectives, and for creating a culture of rigour and accountability around the design, procurement, use and monitoring of AI decision-making systems.

Outline of Topics: After an introduction to the concepts of ethical AI, participants will explore some of the key risks and challenges in making AI systems ethical and how to manage these risks. Topics include:

- *Risks in operating an AI system* - we identify the key risks in data-driven automated decision-systems (including examples of where such systems have led to unintended consequences at a massive scale).

- *AI system governance* - systems that make decisions automatically must still have humans accountable for their actions, and for the design decisions that govern those actions. An aspect of creating ethical AI systems is empowering the right people within the organisation to be responsible for an AI system, and ensuring they have the tools and ability for that responsibility to be meaningful.

- *Building an ethical AI system* - we outline some key stages in the development of an ethical AI system. Each stage involves a different mix of designers, decision makers, stakeholders and engineers. We outline how leaders and decision makers are involved in eliciting and ultimately setting the different objectives (ethical and organisational) of the system and how they are balanced.

- *Managing risks of AI systems* - we discuss strategies that can be used to manage the risks when building or procuring, operating and monitoring AI systems. Course materials include samples of material that can be used at Board level for AI system risk management.

4 Other Gradient Institute Services

As well as providing training, Gradient Institute can provide advice or assistance to:

- help organisations design and implement a new AI system to operate ethically,
- help analyse the ethical intent and impact of an existing system, and find means to improve the system's impact in the future.

This work closely involves both decision-makers and practitioners in an organisation, and as such provides a concrete counterpoint to the training courses.