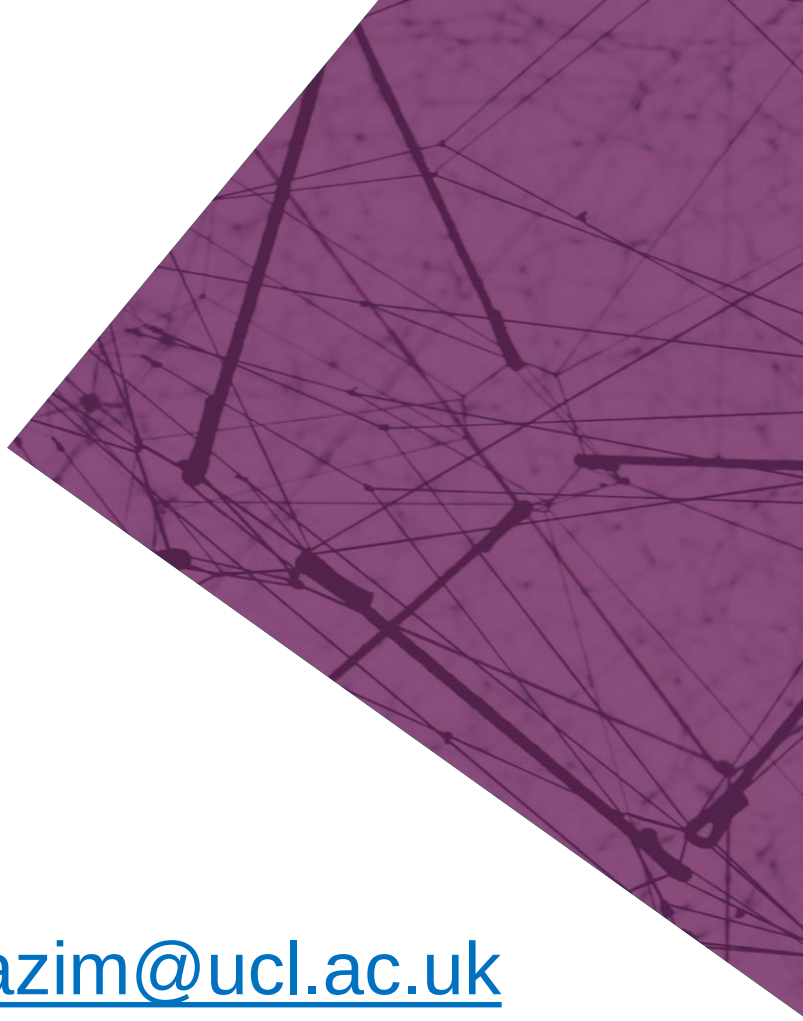
The Alan Turing Institute



# Conceptual Intro. To Fairness for AI Students

Dr Emre Kazim | UCL CS | e.kazim@ucl.ac.uk

The Alan Turing Institute

# What we will cover

- Part 1: Conceptual Approaches to Fairness
- Part 2: Legal Approaches
- Part 3: AI and Fairness
- Part 4: Governance
- Part 5: Case-Study: AI-driven Recruitment
- Part 6: Ethical Dilemmas
- Readings

The Alan Turing Institute

# Part 1: Approaches to Fairness

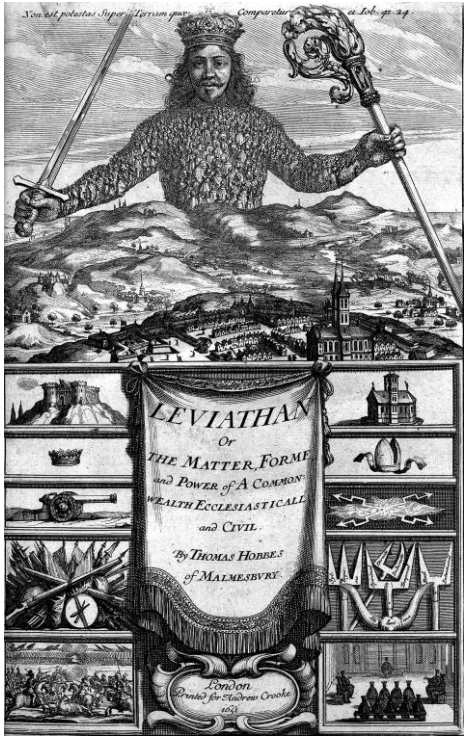
- Human **Dignity**
- Theories of **Fairness**
  - Outcome
  - Opportunity
- Political Economy
- **Context** of Fairness

# Human Dignity

- Foundational idea
- Universal – not context bound
- ‘Agents’: rational and free
- Respect for personhood
- All humans are *essentially* same
- Ergo, all humans are equal
- Humanity

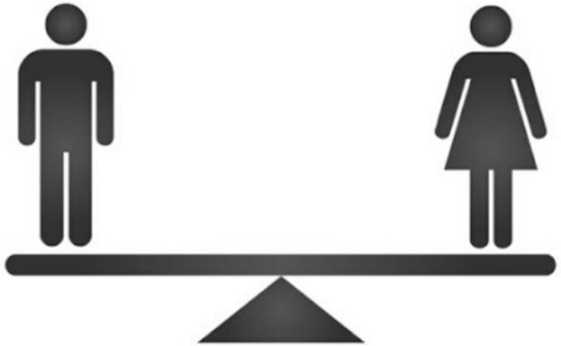
# Theories of Fairness -

- What does it mean to ‘treat’ a person/s fairly?
- Assumes state: collective, resource, mechanisms of exchange, legal recourse -
- Ethical as political <> Political as ethical



# Theories of Fairness - Outcome

- Equality of outcome
  - Employment, uni admission etc…
- What's being made equal?
  - Economic (wealth), social (housing, education, etc.)
- Who's being made equal?
  - Individuals, family, communities
- Redistributive mechanisms
  - Welfare, positive discrimination



# Theories of Fairness – Opportunity

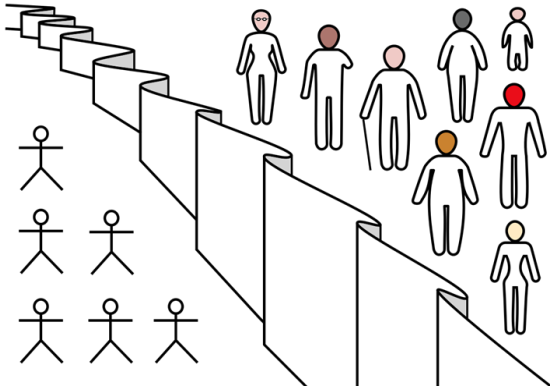
- Equality of Opportunity
  - Procedural
- All treated the same
  - Individual, group
- Processes by which decisions made
  - Criminal justice, recruitment, grading
- Meritocratic
  - Level playing field, equal access

# Theories of Fairness – Political Economy

- Modern state - individual liberty vs collective good
- Equality of Outcome
  - Contravene individual liberty (human dignity)
- Equality of Opportunity
  - ‘Myth’ – people are not ‘born’ equal (in to equal context)
  - human dignity

# Theories of Fairness – Context

- ‘Veil of Ignorance’
  - Thinking *in abstracto*
- Context as King
  - How the world *is*
- History
  - ‘Partnership across generations’
  - Historical injustice
- Scope
  - Local, regional, national, international



The Alan Turing Institute

# Part 2:
# Legal Approaches

- Anti-Discrimination Legislation
  - UK, EU, US
- ‘Protected Characteristics’
- Substantive justice

# Legislation – UK Equality Act (2010)

- Public Sector provision to reduce socio-economic disadv.
- Protected Characteristics
  - Age, Disability, Gender reassignment, Marriage and civil partnership, Race, Religion or belief, Sex, Sexual orientation
- Direct and Indirect discrimination
  - Proxy
- ‘Positive action measures’
  - Act to compensate for those characteristics where reasonable belief of disadvantage

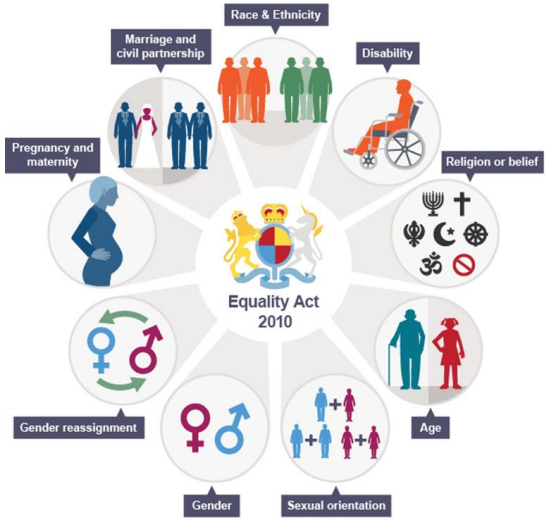



Image from: https://legalresearch.blogs.bris.ac.uk/2016/05/some-thoughts-on-european-and-national-non-discrimination-law-and-brexit/

# Legislation – EU

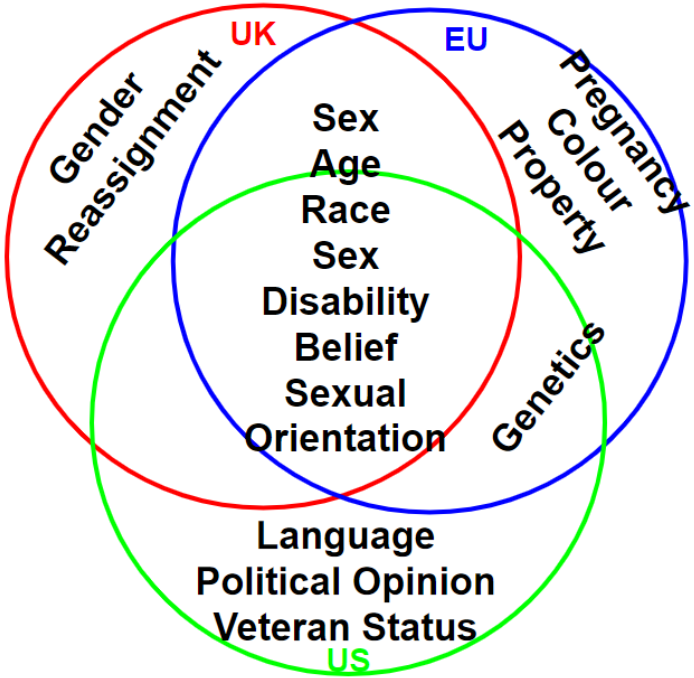
- EU Charter of Fundamental Rights
- Discrimination based on (Article 21)
  - sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation
- Positive action directives
  - Equally qualified candidates of different sexes - Priority given to women (ECJ 1995 Kalanke v Bremen)

# Legislation – USA

- Civil Rights Act (1963, 4, 8, etc…)
- Discrimination based on
  - Race, Religious belief, National origin Sex, Familial status, Disability status, Pregnancy, Age, veteran status, genetic information
- Executive Order 11246 (1965)
  - '…federal contractors […], affirmative action must be taken […] to recruit […] qualified minorities, women, persons with disabilities, and covered veterans. […] include training programs, outreach efforts […]'.

# Themes

- Protected Characteristics
  - Many shared, some differ
  - Problems with definitions
  - Impossible to parallel comply
  - Reasonable pluralism
- Group vs Individual
- Notice: context



The Alan Turing Institute

# Part 3:
# AI and Fairness

- High-profile harm
- AI's fairness problem
- AI ethics
  - Principles
  - Ethical-by-Design
- Which fairness?

# High-Profile Harm






# AI's Fairness Problem

- Human are unfair!
- Amplification risk
- Opportunity?
  - Ex. Judges on Friday

# AI's Fairness Problem – Sources of Bias

- Data
  - Unbalanced – ex. MRI
- Model
  - Classifiers – ex. Nurse (F) / Doctor (M)
- Team
  - Developing
  - Governing/overseeing

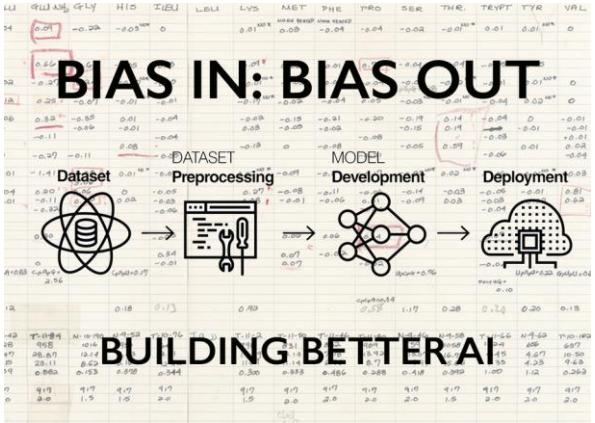



Image from: https://schedule.sxsw.com/2019/events/PP83596

# AI Ethics -

- Rapid dev. New Tech
  - Ubiquity
  - Digital Ethics → AI Ethics
  - Impact: ex. Cambridge Analytica (18’)
- Interdisciplinary
  - Engineering Ethics: values/codes of practice
  - Phil. Technology – doom and gloom
  - STS: value laden
- Utopian/Dystopian
  - Solve all problems/Moral panic

# AI Ethics - Principles

- 100+
  - Industry/Gov./NGO/Civil Society
- Incongruent
  - Transparency vs. Privacy
- Operationalisation
  - Vague/terminological confusion
  - Non-trivial to engineer

| Ethical principle | Number of documents | Included codes |
|---|---|---|
| Transparency | 73/84 | Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing |
| Justice and fairness | 68/84 | Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution |
| Non-maleficence | 60/84 | Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion |
| Responsibility | 60/84 | Responsibility, accountability, liability, acting with integrity |
| Privacy | 47/84 | Privacy, personal or private information |
| Beneficence | 41/84 | Benefits, beneficence, well-being, peace, social good, common good |
| Freedom and autonomy | 34/84 | Freedom, autonomy, consent, choice, self-determination, liberty, empowerment |
| Trust | 28/84 | Trust |
| Sustainability | 14/84 | Sustainability, environment (nature), energy, resources (energy) |
| Dignity | 13/84 | Dignity |
| Solidarity | 6/84 | Solidarity, social security, cohesion |

Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

# AI Ethics – ‘Fairness Principle’

- 23 definitions
  - Irreducible
- C.f ‘messy’ human approaches

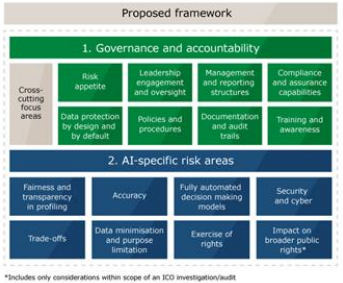
| | Definition | Paper | Citation # | Result |
|---|---|---|---|---|
| 3.1.1 | Group fairness or statistical parity | [12] | 208 | × |
| 3.1.2 | Conditional statistical parity | [11] | 29 | ✓ |
| 3.2.1 | Predictive parity | [10] | 57 | ✓ |
| 3.2.2 | False positive error rate balance | [10] | 57 | × |
| 3.2.3 | False negative error rate balance | [10] | 57 | ✓ |
| 3.2.4 | Equalised odds | [14] | 106 | × |
| 3.2.5 | Conditional use accuracy equality | [8] | 18 | × |
| 3.2.6 | Overall accuracy equality | [8] | 18 | ✓ |
| 3.2.7 | Treatment equality | [8] | 18 | × |
| 3.3.1 | Test-fairness or calibration | [10] | 57 | ✓ |
| 3.3.2 | Well calibration | [16] | 81 | ✓ |
| 3.3.3 | Balance for positive class | [16] | 81 | ✓ |
| 3.3.4 | Balance for negative class | [16] | 81 | × |
| 4.1 | Causal discrimination | [13] | 1 | × |
| 4.2 | Fairness through unawareness | [17] | 14 | ✓ |
| 4.3 | Fairness through awareness | [12] | 208 | × |
| 5.1 | Counterfactual fairness | [17] | 14 | – |
| 5.2 | No unresolved discrimination | [15] | 14 | – |
| 5.3 | No proxy discrimination | [15] | 14 | – |
| 5.4 | Fair inference | [19] | 6 | – |

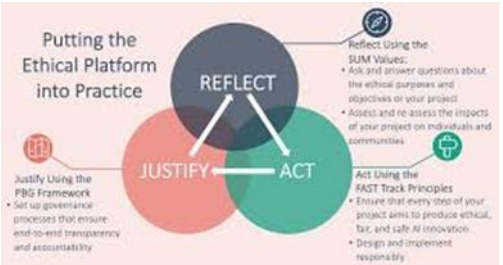
**Table 1: Considered Definitions of Fairness**

Verma, Ruben, 2018 ACM/IEEE International Workshop on Software Fairness
https://fairware.cs.umass.edu/papers/Verma.pdf

# AI Ethics – Ethical-by-Design

- Ex-ante risk mitigation
- Co-design
  - Ethicists, lawyers
- Build to be audited
  - Explainable
  - Justified design
- Practical manuals
  - IEEE; ATI; ICO




# AI Ethics – Which Fairness?

- Mutually exclusive definitions
  - Pick one?
  - Respect reasonable pluralism → Prioritise on context
- Discernment and fairness
  - Boundary conditions (ex. UK Eq. Act 2010)
- Who Discerns?
  - Human vs machine
  - Authorised/democratic legitimacy
  - Competency (judges don't get it!)

The Alan Turing Institute

# Part 4: Governance

- EU AI Regulation
  - Risk-Management System
- Accountability (Non-Technical)
- Performance Assessments
  - Fairness Metrics

# Governance – ex. EU AI Reg.

- Proposed EU AI Regulation (April 2021)
- Risk Approach
  - Unacceptable - Prohibited
  - Contexts/use-cases - High-Risk
  - Minimal/low risk - Transparency
- Legal requirements for Risk-management
  - Data governance
  - Documentation
    - Record keeping
    - Technical (accuracy/robustness/fairness)
  - Transparency for users
  - Human-oversight
- Abstract

EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

**REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL**

**LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS**

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

# Governance – Accountability

- Articulate ethical principles
  - Position on fairness
- Taxonomy of accountability roles
  - Compliance, data science
  - Hierarchies
- Monitoring
  - Reporting mechanism
  - Accessible monitoring (interface?)
  - Appropriate human-oversight
- Impact assessments
  - Proportionality
  - Social impact, impact on people

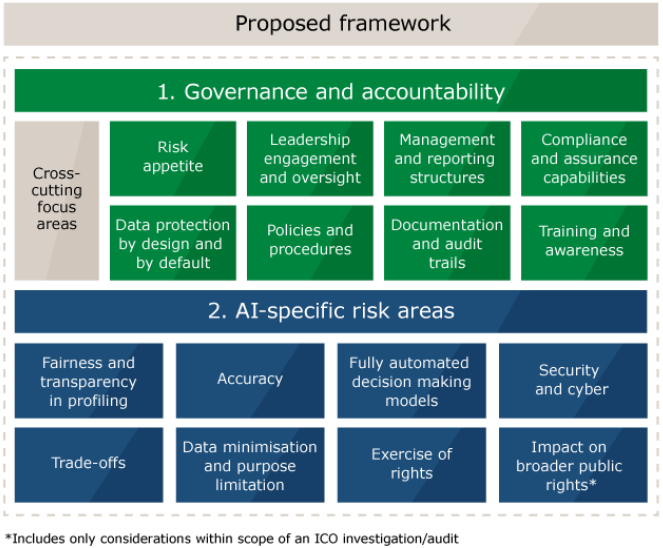



Image from: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

# Governance – Performance Assessment

- Technical assessments
  - Stress testing
- Data
- Model

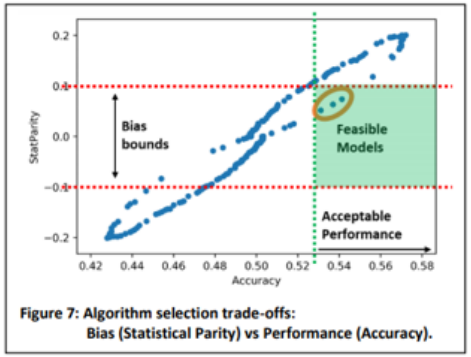



Figure 7: Algorithm selection trade-offs: Bias (Statistical Parity) vs Performance (Accuracy).

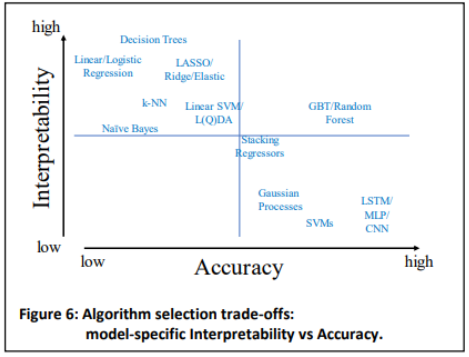
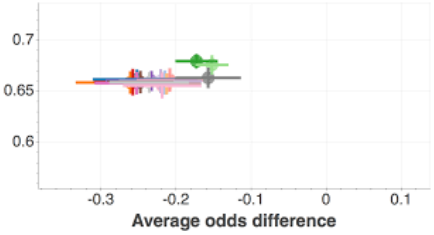


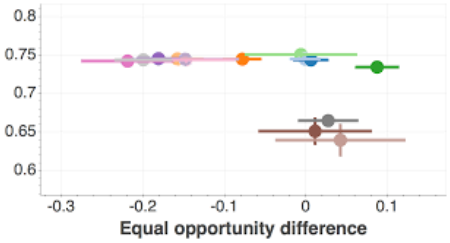

Figure 6: Algorithm selection trade-offs: model-specific Interpretability vs Accuracy.

Koshiyama, Adriano, et al. Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms (January 2021). Available at SSRN: https://ssrn.com/abstract=3778998 or http://dx.doi.org/10.2139/ssrn.3778998

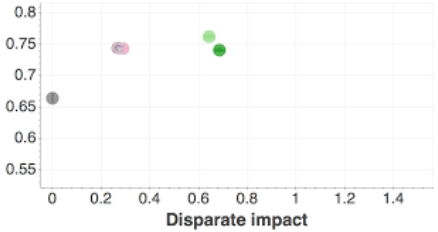
# Governance – Fairness Metrics

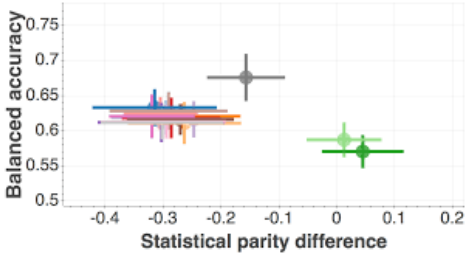
- Equality of Opportunity
  - Equal Opportunity Difference
  - Average Odds Difference

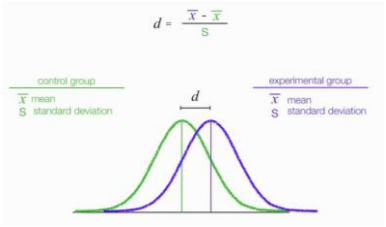



- Equality of Outcome
  - Disparate Impact
  - Statistical Parity
  - Cohen's D2-SD Rule



The Alan Turing Institute

# Part 5: Case-Study – Recruitment

- AI-Driven Recruitment Industry
- Ethics
- How is the system being used?
- What is being assessed?
- Audit

# AI-Driven Recruitment Industry

- Aggregating appropriate candidates
- Interviewing through chatbot
- Video interview assessment
- CV-analysis
- Also used to automate HR processes
  - Posting job listings
  - Matching to job roles
  - Performance measurement
  - Salaries





# Ethics

- Impact life prospects
  - Opportunities/Wages
  - Experience/Networking
- Reflect social bias
  - Mirroring effect/Hiring oneself
  - Signalling (accent, clothing, etc.)
- Assessment Bias
  - IQ, psychometrics - raw vs trained ability
  - Oxbridge interview - accustom vs alien
- Human Oversight
  - Meaningfully assess recommendations
  - Provide explanation

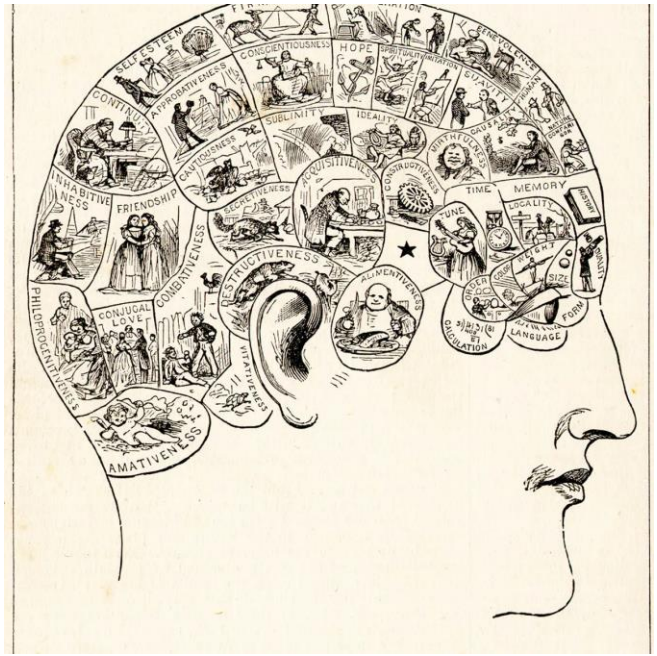
# How is the system being used?

- Where in the pipeline?
  - Mass screening – grad program
  - Later stages
- Reflect social bias

Image from: https://www.altamirahrm.com/en/blog/recruiting-funnel

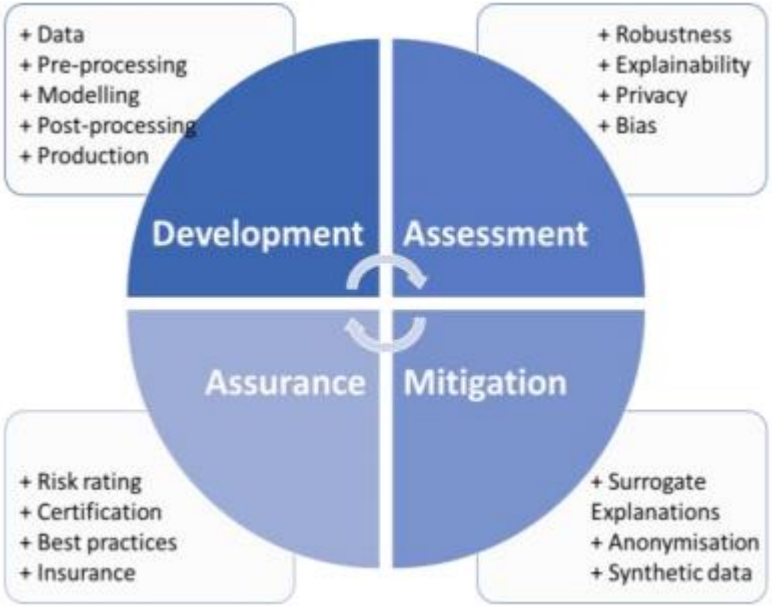
# What is being assessed?

- Sentiment analysis
- Facial Expressions
- Voice
- Text
  - Written answers
  - CV
- Sentiment analysis as Phrenology?



# Audit

- Mandatory bias audit (NYC)
  - Fairness metrics
  - Third-party inspection ready
- Assurance and Certification



The Alan Turing Institute

# Part 6:
# Ethical Dilemmas

- Monitoring at work
- Recidivism Calculation
- Facial recognition in public square

# Monitoring at Work

- Company decides to employ AI to analyse behaviour at work through visual/audial monitoring
  - Productivity and behaviour
  - Reward ‘good’ behaviour and punish ‘bad’
- Covid-19 strikes - work from home; employees monitored through tracking keyboard/mouse, email, and camera (same reward structure)
  - List 3 fairness and bias concerns; explain why these are concerns and how they can be addressed
  - Discuss ethical significance of shift to home work in terms of fairness

Image from: https://www.lawyer-monthly.com/2019/09/surveillance-at-work-how-far-is-too-far/

# Recidivism Calculation

- Recidivism - likelihood of reoffending. Informs sentencing. Systems have been shown to be bias towards certain groups
- Assuming bias is due to data bias and that this could be addressed – in terms of fairness, is there an ethical issues with using a de-biased system? Give reasons for and against

# Facial Recognition in Public Square

- FR can be used to identify a person ‘live’. Police are deciding how/where to use – systems have been shown to be inaccurate (bias) especially to particular demographics –
- Is it fair to use these systems to identify known criminals?
- Is it fair to use in airport passport control?
- Is it fair to use in the context major sporting events?

The Alan Turing Institute

# Readings

- Readings have been provided for the respective parts of this introduction
- These readings should form the point of departure for further reading

# Readings – Part 1

- A History of Human Dignity - https://blogs.lse.ac.uk/theforum/a-history-of-human-dignity/ Remy Debes. February 5th, 2018
- EU Charter of Fundamental Rights (article 1) https://fra.europa.eu/en/eu-charter/article/1-human-dignity
- Arneson, R. (2002). Equality of opportunity. https://plato.stanford.edu/entries/equal-opportunity/
- Hanna, R., & Kazim, E. (2021). Philosophical foundations for digital ethics and AI Ethics: a dignitarian approach. AI and Ethics, 1-19. https://link.springer.com/article/10.1007/s43681-021-00040-9

# Readings – Part 2

- UK Equality Act 2010 https://www.legislation.gov.uk/ukpga/2010/15/contents
- EU Charter of Fundamental Rights (article 21) https://fra.europa.eu/en/eu-charter/article/21-non-discrimination#:~:text=Next%20article-,Article%2021%20%2D%20Non%2Ddiscrimination,sexual%20orientation%20shall%20be%20prohibited.
- Civil Rights Act United States [1964] https://www.britannica.com/event/Civil-Rights-Act-United-States-1964
- Fullinwider, R. (2001). Affirmative action. https://plato.stanford.edu/entries/affirmative-action/
- List of US AI regulation-

# Readings – Part 3


- Kazim, E. & Koshiyama, A. ‘A High-Level Overview of AI Ethics’ (May 24, 2020). http://dx.doi.org/10.2139/ssrn.3609292
- Kazim, E., et al. ‘Automation and Fairness: Assessing the Automation of Fairness in Cases of Reasonable Pluralism and Considering the Blackbox of Human Judgment’ (September 24, 2020). http://dx.doi.org/10.2139/ssrn.3698404
- CDEI. (2020). Review into bias in algorithmic Decision-making, . https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/939109/CDEI_review_into_bias_in_algorithmic_decision-making.pdf
- European Commission (2020). Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation

# Readings – Part 4

- Koshiyama, A., et al., (2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3778998
- Examining the Black Box: Tools for assessing algorithmic systems. (2020). Ada-Lovelace Institute & DataKind UK. https://www.adalovelaceinstitute.org/examining-the-black-boxtools-for-assessing-algorithmic-systems/.
- Andrade, N. N. G.,, & Kontschieder, V.. “AI Impact Assessment: A Policy Prototyping Experiment” (2021), https://d32j3j47emgb6f.cloudfront.net/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf

# Readings – Part 5


- Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: implications for recruitment. Strategic HR Review. https://www.emerald.com/insight/content/doi/10.1108/SHR-07-2018-0051/full/html
- Davenport, T. H. (2018). From analytics to artificial intelligence. Journal of Business Analytics, 1(2), 73-80. https://www.tandfonline.com/doi/full/10.1080/2573234X.2018.1543535
- Hadjimichael, D., & Tsoukas, H. (2019). Toward a better understanding of tacit knowledge in organizations: Taking stock and moving forward. Academy of Management Annals, 13(2), 672-703. https://doi.org/10.5465/annals.2017.0084