## Power, trust, and machines

Goals of trust :Trust is not the end goal

Trust between people: Predictability, enable “collaboration”

Human-machine trust: Predictability, enable collaboration

Trust between human: view from sociology:

A trusts B if:

1.A believes that B will act in A’s best interests

2.A accepts vulnerability to B’s actions

3.so that A can anticipate the impact of B’s actions, enabling collaboration.

Trust between human-machine (H: person .M: machine)

1.H believes that M will act in H’s best interests

2.H accepts vulnerability to M’s actions

3.so that H can anticipate the impact of M’s decisions, enabling collaboration

Ethically, trust is only desirable if it is warranted

Distrust is desirable if it is warranted

Distrust VS lack of Trust

1.H distrusts M if:

H believes that M will act against H’s best interests.

2.Lack of trust :(absence of trust)

1.H does not believe that M will act in H’s best interests.

2.H does not accept vulnerability to M’s actions

CONTRACTUAL TRUST(契约信任): THE VIEW FROM SOCIOLOGY

1. social or normative contract, as well as legal.
2. Contractual trust=humans trusting humans to fulfill a contract in a particular context

Contracts for Artificial Intelligence:

图形用户界面

描述已自动生成

TRUST AND TRUSTWORTHINESS

A machine is trustworthy if:It can fulfill its set of contracts

Trust does not imply trustworthiness

Trust does not imply trustworthiness

有根据的信任和不信任

图形用户界面

中度可信度描述已自动生成

Use, misuse , disuse and abuse: unwarranted trust and distrust.

Factors that determine use of automation:

Mental workload/ cognitive overload / trust

Misuse of automation(误用)（疏忽无知造成的）

Definition: using automation when it not should be used

Cause: unwarranted trust / over-reliance on automation/ decision biases and automation biases /machine monitoring errors

Impacts: complacency(自满)

Disuse of automation;(废弃)

Definition: Not using automation when it should be used

Cause:

unwarranted trust(没有根据的不信任)

human monitoring errors(law false alarm rate)

human bias

Impacts: disabling/ignoring alarms

Abuse of automation（滥用、有意的）

Definition: Deploying automation when it should not be

Cause:

unwarranted distrust(对人没根据的不信任)(from designer)

distrust in human operators

automation bias

arrogance

Impacts: mismatch in human-automation interface

图形用户界面, 文本, 应用程序

描述已自动生成

图形用户界面

描述已自动生成Power, trust and machines

Power: the ability to control our circumstances

Trust ≠ Ethics ≠ Power

Ethical issues emerge from (real or perceived) power imbalances between groups with

different interests.

History of AI

i)golden age of AI

ii)The 1st AI winter: (Outcomes failed to live up to the hype)

1. scalability
2. commonsense knowledge
3. perceptron limitation
4. MORAVEC’S PARADOX

Result: funding dried up; global interest in AI died down; criticism from philosophers and cognitive scientists

iii)The knowledge era:

iv) the 2nd AI winter:

1.scalability 2.maintence 3.the qualification problem 4. MORAVEC’S PARADOX

Result:

1. FUNDING DRIED UP (DARPA DECLARED AI WAS `NOT THE NEXT WAVE’)

2.GLOBAL INTEREST IN AI DIED DOWN

3.AI COMPANIES WENT BANKRUPT

vi)The AI revival

WHAT ARE SOME OF THE RISKS?

By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it.

Effects of under-representation:

Lack of diversity means lack of: Privacy safety fairness transparency functionality accessibility&inclusion

手机屏幕的截图

描述已自动生成

Diversity in teams leads to diversity of input \_\_\_\_good for business!

Utilitarianism and deontology

principles in AI ethics:

Fairness

Safety

Accountability

Transparency

Benefit

Explainability

Privacy

What is ethics?

Socrates: How should one live? Better to suffer evil than to do it

Ludwig Wittgenstein: “Ethics is the enquiry into what is valuable, or into what is really

important, or into the meaning of life, or into what makes life worth living, or into the right

way of living.

Ethics about values, morals, good and bad, right and wrong

1. How should I act? What standards of behavior should I adopt?

2. What sort of person should I be?

3. What sort of professional computer scientist should I be?

Self and others

Nihilism: no standards (do what you want- no ought) but self-centred standards

Egoism: I ought to do only does the most good for me

Utilitarianism

Consequentialism: consequences alone determine action’s rightness

Greatest-Happiness Principle: “actions are right in proportion as they tend to promote

happiness, wrong as they tend to produce the reverse of happiness”

Greatest happiness of the greatest number of people

Principle of utility: Right = maximise total net happiness/wellbeing

Hedonism: “Two masters”: pleasure & pain

Bentham: pain/suffering is bad. Intensity, duration. all pleasures are equal: pushpin as good as poetry.

Mill: higher and lower pleasures: poetry better than pushpin. Socrates unsatisfied>satisfied fool

Strengths:

1. surely happiness/ wellbeing matter.
2. Surely consequences matter
3. Simple ,clear decision procedure :principle of utility
4. Rational (cf: accepting authority)
5. Equality : all count for 1:no class. Race, gender, intelligence, etc. favouritism

Deontology:

Non-teleological(ends); not purely consequentialist;

rules of principles;

we learn rules and principle from childhood

can refine and alter rules via reflection and argument.

文本

中度可信度描述已自动生成

WD Ross:

1. Fidelity: keep promises and be honest and truthful
2. Reparation: make amends when we have wronged someone
3. Gratitude: be grateful when others benefit us and try to return favor.
4. Non-injury(non-maleficence(非恶意)):refrain from harming others physically or psycholohically.
5. Beneficence: be kind and try to improve others’ health, wisdom, security, happiness, and well-being.
6. Self-improvement. Strive to improve our own health, wisdom, security, happiness, and well-being.
7. Justice. Be fair and distribute benefits and burdens equably and evenly.

D attacks U

**For some deontologists:** Consequences can matter – e.g. generosity requires

calculating benefit. But – more to ethics than calculating consequences

**Maximising ethic too demanding** – give up much of own wellbeing for strangers.

Singer - give up high percentage of income

**Not as helpful a decision-making procedure as U think** – difficult to impossible

calculation

**Fairness –** although each person’s similar interests counts equally, maximizing

wellbeing can cause apparent injustice

prima facie vs absolute rules.

Prima facie rules:

Rule applies presumptively

rules can conflict: need judgement to resolve (break promise to save a life)

some rules win out ,others are overridden

absolute rules: unconditional

don’t have exceptions; don’t yield to other rules

greatest protagonist: (最大的主角)

Kant’s ethics(special kind of deontology)

Reading：

Kant gave closer attention to the problem of how his supreme formal principle of morality can provide guidance in concrete situations. One of his examples is as follows. Suppose that a person plans to get some money by promising to pay it back, though he has no intention of keeping his promise. The maxim of such an action might be: “Make false promises when it suits you to do so.” Could such a maxim be a universal law? Of course not. The maxim is self-defeating, because if promises were so easily broken, no one would rely on them, and the practice of making promises would cease. For this reason, the moral law would not allow one to carry out such a plan.

和utilitarian比的优势：Modern Kantians hold that it does, because they interpret it as denying the legitimacy of sacrificing the rights of one human being in order to benefit others.

Kant的局限性：Another possibility would be to formulate the maxim of the action with sufficient precision to define the circumstances under which it would be permissible to tell lies—e.g., perhaps there could be a universal law that permitted lying to people who intend to commit murder. Kant did not explore such solutions, however.

Absolute duties

Morality is rational

Right =a ‘good will’ acting for right reasons; acting for duty’s sake

But rationality is opposed to consequentialism

Categorical imperative-first

Actions must be universalizable

We act on rules in ethics

But moral rules don’t just apply to you

Can’t make exception “Act only according to that maxim (rule) whereby you can at the same time will that it should become a universal law”

Rule: it’s OK for me to lie

E.g.: Rule: it’s OK for me to lie

• This means: it’s OK for anyone to lie

• But if everyone lies when it suits them: truth collapses > lying becomes

impossible

• Hence: lying is irrational

• Same for promise-breaking

Categorical imperative-second

All rational beings can grasp and follow the moral law. They have autonomy

Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end”——人不是手段而是目的本身

（杀手问你藏没藏受害人——尊重杀手的自主性——告诉他藏了）

举一个具体的例子，如果你在办公室里，你的同事只是因为他们可以帮助你完成工作或者获得晋升而被你看待和利用，那么你是把他们作为手段来看待的。相反，如果你尊重并认可他们作为独立的、具有自我价值和目的的人，即使他们无法为你提供任何帮助，那么你就是把他们作为"目的本身"来看待。

Modern notion of autonomy：

Autonomy is the ability to think for ourselves, plan our lives, act on our values

To respect autonomy, need people’s informed consent (e.g. collecting private

information about them)

We should aim to:

• Respect the autonomy of others

• Try to understand people’s values and beliefs and get their consent

• Respect control over personal information

• Remember both powerful and weak have autonomy

• Be honest with people, including when things go wrong (deceiving people

disrespects their autonomy)

U and rules:

Consequences matter more than rules

But: rules matter if they affect consequences !( social rule against punishing innocent-good U rules?)

social rule against punishing innocent---good U rule.

Some rules, laws basic rights are important (don’t kill don’t torture )

But Must be changed if not best consequences!

U: ’morality made for people ,not people for morality’.

Virtue ethics:（ethic of character 伦理品格、一种遗留下来的宗教规则）

1. Being a kind of person -- praiseworthy, warrants respect, emulation.(成为一个值得表扬、尊敬、模仿的人)
2. Ethics:aims not just at right action, but good way of being（目标不只是正确的行动，而是良好的活人方式）
3. kant: shopkeeper not ripping-off customers: pure duty/good will vs.caring about
4. Deontology/consequentialism:ignore centrality of feeling/emotion/attitude
5. Aretaic ethics: excellence of character(cf. intellect )

Aristotle: Virutes

Dispositions or traits

Cultivated over time

Habit-settled dispositions(从小养成的：奖励、惩罚、做榜样)

Kant ‘good will’/follow duty(not feeling /inclination)

Aristotle:Right attitudes /motivations/emotions

Example:visit sick friend(Kant:处于道义伦理义务规则我该去；aristotle:出于我对朋友的同情、个人的情感)

Aristotle: Doctrine of Golden Mean

Right feeling/action/thoughts

Character traits: on a spectrum

Virtues: the sweet spot

Excess and deficiency =vice

Affected by context

Example: lazy student steals your hard work (Indifference---Righteous indignation---Spite)

Trust in AI: Gullible – Right trust ---cynical

Affected by situation /context eg. AI in Netflix vs. predictive policing

Also: people might disagree over how to apply a virtue E.g. Also people might disagree over how to apply a virtue e.g. some will say that AI policing can be just, others will say a just person will never support it!

You need to use judgement and reasons to explain what is virtuous and why

Criticism: VE doesn’t tell us what to do ,only how to be

Phronesis:

1.Practical wisdom(virtue)

2. Apply moral ideas/rules contextually

3.VE versatile, flexible, sensitive to situations

What a virtuous person would do?

How do we know that?

We first learn from education, training, school, parents.

Aristotle: young people and ethics

And we begin to look to moral exemplars

Aristotle: Virtues are constitutive of good life

Ben-pretends to be virtuous, but is greedy, callous,unjust,dishonest,unfaithful and has loads of fun

Lei-virtuous and good ,but is poor and struggles with chronic illness.

Without virtues: your life goes badly, even if rich, famous, successful

Ethical relationships part of good life

Aristotle: happiness=flourishing=eudaimonia(幸福感)

Socates苏格拉底（遭受恶魔比做恶魔要好）

Moral rules follow from this.

Virtuous exemplars:  
the female stood up to big tech: called for fairness, transparency, diversity in AI development

What virtues might they have? Just? Honest? Courageous?

Technological convergence: discrete technologies merging and magnifying their scope and power(离散技术融合并扩大了它们的范围和力量)

Technology:

Growing techno-social opacity

Unpredictable consequences

Future individuals-human, nonhuman

Has trans-cultural resonance

Criticism of VE(from U and D):

1. Virtuous people still need to act on/ for reasons:Reasons include rules, duties, principles, concern for consequences/utility(yes, but the virtues help us decide those things)
2. Too vague: no specific guidance on moral dilemmas (Think: what virtues? What exemplars? How to apply virtues? What are the circumstances? Consider the details.)
3. Duty or consequences are primary(责任和结果是主要的): character only valuable instrumentally（品格仅在工具上有价值）Character necessary for having good judgement about duties/consequences（要做出关于指责和后果的良好判断，一个人必须具备良好的品格）

**Care ethics**

促使我们关注社会中那些边缘化的，弱势群体尊重他们的autonomy和人格，从而以此achieving a full range of human flourishing

A Feminist ethics

Moral development of children

male: steal drug! Because:

Human life more basic value than property

Fairness

Female:

Relationships and caring

Talk to chemist and emphasize dire situation of wife-empathy

Stealing>jail>wife left alone

Possibly same conclusion, but

Masculine vs feminine perspective

Masculine: ‘ethic of justice’ and ‘rights’

Masculine higher-based in detached think(独立思考) ,rules ,principles, consequences.

Carol Gilligan

In a different voice(1982)

Women’s voice sidelined(被边缘化); seen as atypical（异常）

rationalism, principles, severe impartiality, lack of emotion(理想主义、原则性、严格的公正性，缺乏情感)

Very critical of U and D!

Relational and interpersonal- emotions and actions

Attending, listening, loving, feeling, taking, responsibility

Recognizing vulnerability, powerlessness

Protecting relationships

Joan Tronto

Care: a species of activity that includes everything we do to maintain, contain, and repair our ‘world’ so that we can live in it as well as possible. That world includes our bodies, ourselves and our environment. Sustain our own and others’ basic needs, alleviate pain/suffering, enable mutual flourishing

Care: care begins in close relations eg. Family

Often face-to-face and embodied

Relationship: carer and cared-for-interdependence

Mutual vulnerability: all might suddenly need care

Self-sufficiency myth (cf. Kant’s autonomous person)

Essential-healthy society

Part of VE: but care central to good character/action

Joan Tronto:

Care involves:

1. attentiveness: inclination to become aware of need.( a proclivity to become aware of need)
2. Responsibility: willingness to respond to need(responsibility, a willingness to respond and take care of need)
3. Competence: skill providing good successful care(competence, the skill of providing good and successful care;)
4. Responsiveness: consideration of the position of others as they see it[empathy ], recognition of potential for abuse in care, reviewing and adjusting

Various emotional responses, attitudes, inclinations ( e.g. empathy, compassion)( responsiveness, consideration of the position of others as they see it and recognition of the potential for abuse in care)

Intelligence and judgement

Care 和service的区别：care is understood as a virtue or motive

Bubeck thus distinguishes care from “service”, by stipulating that “care” involves meeting the needs for others who cannot meet their needs themselves, whereas “service” involves meeting the needs of individuals who are capable of self-care

Eva Kittay

1. Emphasizes human dependency
2. Disabled, impaired ,disadvantaged people(e.g. sesha: can’t walk , talk, or read---Help to flourish as best she can)

Criticisms of CE:

Too vague(like VE)

CE: can also use principles ,rules that emerge from caring. Some vaguenesss is important –ethics is contextual.

Romanticises motherhood? slave morality? Depicts carers(often women) as too self-sacrificing?

CE:care should be mutual, not women only; men are carers; carers need care too

Criticisms

Feminine perspective a stereotype?

CE: many women stress impartial justice; men who stress care; not saying that care is completely gendered.

Parochial to own people/circle? (狭隘) need impartial justice?to strangers, third-world, intersectional oppressions (e.g. trans black people)

Parochial to Own People/Circle: 在伦理学和社会学中，"parochial"常被用来描述一个过于狭窄或限定的视角，只关注自己的人群或社区，忽视或无法理解其他人群或社区的经验和需求。这种视角可能导致偏见和不公正的决策。

Need for Impartial Justice: 为了应对这种偏狭视角，我们需要实施公正的原则，这就是公正的不偏不倚。公正应当是公平的，不偏袒任何特定的人群，不论是你自己的人群还是其他人群。

To Strangers, Third-World, Intersectional Oppressions (e.g., trans black people): 当我们考虑公正时，我们需要特别注意那些经常被忽视或被压迫的人群，如陌生人、发展中国家的人民，以及处于交叉压迫的人群。例如，黑人跨性别者可能同时面临种族和性别的压迫，这种交叉压迫使他们更容易受到不公正的对待。

CE: True, care often stronger for those we are in special relations with(对与我们有特殊关系的人往往关心更强烈)

But: stresses vulnerability and oppression(但强调脆弱与压迫)

And: justice an extension of caring(公平是caring 的一种延伸)

Care can lead to and be involved in justice (e.g. view others as children, parents, brothers, sisters in need of care)

Can have caring relations beyond one’s immediate circle.

Principlism: 4+1

Derived from medical ethics

1. non-maleficence-do no harm----predict harm, avoid causing harm, minimize harm, short and long term
2. beneficence-do good: anticipate good outcomes, short and long term.
3. Respect autonomy: respect people’s values, choices, life plans: understand what other’s value, don’t override their choices ,be honest.
4. Justice: fainess-----distribute benefits and harms fairly, fair processes,don’t unfairly discriminate

4+1. Explicability: transparency and accountability(Floridi\*)----complements the 4 principles. Ensure those potentially impacted have sufficient understanding of the AI and that relevant people are held to account.

Principles: need to be balanced against on another ;all are ‘equal’

Fairness, Accountability, principlism

Compass-----Race Disparate impact

Justice-fairness often used interchangeably(公平正义经常相互交换)

High stakes decisions-AI (高风险决策)\_判刑、战争、找工作、疾病诊断、保险

Equal regardless of race, religion,class,sex,gender,sexual orientation,etc.(equal treatment/consideration/respect)

U-all similar interests are counted equal in calculus

D-duty of justice

Kant-Dignity equal;2nd C.I: always ends ,never merely means

VE-justice :praiseworthy trait

CE-justice in care

Distributive justice:

Resources, opportunities

Need, merit contracts etc.

Negative justice:

Should everyone be given the same career opportunities regardless of talent?

Income?

Should we help people suffer bad luck?

Should we favour people who are morally responsible rather than selfish?

More resources/opportunities/advantages to disadvantaged and historically oppressed groups.

Positive/reverse discrimination:（反歧视）

Affirmative action(平权运动)

More resources/opportunities/advantages to disadvantaged and historically oppressed groups.

Procedural justice/fairness(程序上的公平正义)

Fair procedure or process allocate benefits or harm(公平的过程分配利害)

许多人只关心过程是否公平，不关心实际的结果

Pure procedural justice: no question of need/merit（只是纯关心过程，不需要关心需求和价值等因素）

Retributive justice:报应的正义公平性

Impose penalty due to wrongdoing

Based on actual guilt and fair procedure(trail)(基于实际的罪行和公平的程序——审判)

Reparative justice: remediation for unfair treatment(为不公平的治疗做补偿)

Accountability: Assuming accountability(responsibility)

Eg.holding myself to fair procedures(让自己遵守公平的程序)

Being held accountable(responsible):by external pressures and mechanisms(外部的压力和机制)

e.g. law, profession, codes of practice ,clloeagues, elections

who is accountable? Eg.researcher \engineers\organizations(private and public)

Algorithmic fairness(算法的公平性)

ML algorithms can have embedded bias-unfair

E.g.(discriminate against groups unfairly 不公平的歧视)

Race,gender(种族、性别)

Either explicitly or by prox(明确的或者不明确的)

Technical solutions to minimize unfairness

E.g. change inputs/improve processing of dataset/change weight of false

Mathematical measures:

1. overall accuracy equality

2. statistical parity

3. conditional procedure accuracy equality

4. conditional use accuracy equality

5. treatment equality

6. total fairness (1-5 achieved)

The inherent tradeoff :between fairness and performance

Increased group fairness----decreased accuracy of recidivism prediction for bail

Decrease false positives(降低了假阳性率：没病的人被错误识别为病人): (defendants falsely scored as high risk) but increase false negatives(很多病人没识别出来)

Utilitarianism:fair/just=maximizes net wellbeing,even if some individual’smust be made worse off than others. Everyone’s similar are still considered equal.

Kant’s deontology: recognize human dignity and respect autonomy. (treat autonomous agents always as ends, never merely as mean)

Virtue ethics: consider what a fair and just person would do

Ethics of care: consider special relationships and roles and responsibilities that flow from them. Consider impacts on the most vulnerable

All these frameworks recognize basic human equality

They may have different/similar views on algorithms that: (不同伦理对与算法的观点不同)

Reinforce existing disadvantage（强化了现有的不平等） e.g. increasing policing for some groups

Overlook past oppression e.g. that affect algorithmic prediction that results in biases against disadvantaged groups or don’t positively discriminate to help disadvantaged groups

算法忽略了过去的压迫：算法没有帮助到弱势群体反而造成一种歧视

Pak-hang Wong (reading)

Wong: “ensure decisions are morally and politically acceptable to those affected by algorithms through inclusion and accommodation of their views and voices”

AI may necessarily create winner and losers-harms and benefits for different people.

Determining what is fair in high stakes AI is not purely a technical task, but an ethical one.

But : how to ensure this determination is itself fair?

Wong: procedural justice

What mechanism is fairest for holding AI designers and owner accountable?

E.g. panel of AI ethics experts?

Political Mechanism

Accountability for reasona**bleness (AFR): AFR a**ssumes no totally final and ‘right’ answer: answers emerge through open, democratic, good-faith dialogue and reason-giving involving stakeholders

Not just developers and researchers determining what is fair AI

4 conditions for AFR:

1. **Publicity condition:** Decisions that establish priorities in meeting [algorithmic fairness] and their rationales must be publicly accessible.
2. **Full Acceptability condition:** Justify why the chosen parameters and impacts are relevant. Could those impacted accept these reasons? E.g. allowing ‘disparate impact’ on historically disadvantaged black people while ‘avoiding disparate treatment’? Using education level or being the victim of crime - that negatively affects certain racial groups more?（可接受的参数和相关影响）(The rationales for priority-setting decisions should aim to provide a reasonable explanation of why the priorities selected are thought the best way to progressively realize [the value the algorithm aims to provide] or the best way to meet [claims] of the defined population under reasonable [...] constraints. Specifically, a rationale will be Breasonable^ if it appeals to evidence, reasons, and principles that are accepted as relevant by (fair minded) people who are disposed to finding mutually justifiable terms of cooperation. An obvious device for testing the relevance of reasons is to include a broad range of stakeholders affected by these decisions so that the deliberation considers the full range of considerations people think are relevant to setting priorities.)
3. **Revision and appeals condition: Have ongoing (not one-off)** Mechanism for those impacted to contest those reasons e.g. vulnerable groups, representatives of broader society（保留复审申诉环节）(: There must be mechanisms for challenge and dispute resolution regarding priority-setting decisions and, more broadly, opportunities for revision and improvement of policies in light of new evidence or arguments.
4. **Regulative condition:** Media put spotlight on COMPAS, but no stronger regulation or enforcement(监管)( There is public regulation of the process to ensure that conditions (1)-(3) are met. (强监管))

Other accountability mechanisms? Audits\committees\ ‘Turing’ stamps - suitable body(合适的评级实体机构)/ Open-source software – ‘crowd’（开源软件）

Transparency

decision&processes :overarching processes and decision-making of an

algorithmic system

We include the technical considerations about interpretability of the algorithms themselves, counterfactual analysis, etc under the banner of explainability

Institutional transparency: it requires clarity in the procurement, implementation and technical mechanisms associated with automated decision-making systems.This type of transparency is useful for keeping track of impacts of decision systems over time,and achieving some public disclosure on their purpose, reach ,policies and techniques.

requirement not only for the algorithms, but also the contexts surrounding their implementation.

Automated decision-making system(ADM):safeguard built into the system are only asking relevant questions, telling customers why questions are being asked(which makes the decision-making process more transparent) and recording and explaining to a customer the reason for decision”

“Expert systems’ ability to provide an audit trail of the administrative decision-making processes they are involved in is important to the administrative law values of transparency, fairness and efficiency.

A good system of internal review is one which is transparent in process and affords a quick, inexpensive and independent review of decisions. Such a system is beneficial both to applicants and agencies

A number of provision(条款) in GDPR seek to promote a high degree of transparency in the processing of personal data.In general these provisions require data controllers to provide data subjects with information about the processing of their personal data and to do so in a concise,transparent,intelligible and easily accessible form ,using clear and plain language.

Where personal data are obtained from the data subject,article requires data controllers to provide data subjects with information about ‘the existence of ADM’,including profiling…and meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. The purpose of such information provision is said to be ‘to ensure fair and transparent processing’

Transparent is far more than code and data-------transparency is required in planning, implementation ,auditing…concerns the data ,design,testing ,deployment.also consider legal aspeccts&philosophical concepts in the broader sense:incl.fairness, recourse.

Let’s say I’ve been shown an ad (sponsored post) on Facebook.

As a consumer, I want to know WHY.

FB gives me some reasons.

... but also “more factors not listed”.

... and directs me to a huge page with many explanations, but all “mays” (we may, advertisers may...)

... and has a long TOS explaining the many ways

At present, the platform’s transparency measures offer “nominal transparency”, with no real regard for whether people actually can easily access, read and gain insight into the information held about them and whether this transparency in name foster users’ autonomy.

Aiming for effective transparency-which demonstrably enables users to understand what platforms do with their data and what users’ choices imply, and to then translate this knowledge into measurable behaviour- is an important step towards more acceptable business practices and towards regaining some of the lost autonomy for users(by prompting people to adjust their privacy settings)

Issues at all stages: crime stats; data dossiers; personal/cultural bias; data entry/analysis; tech complexity; financial/IP interests; auditing; metrics...

Fix transparent issue: auditing, public release of metrics, training

generative AI

One key issue is the lack of transparency in the data sets used to train generative AI systems.(plagiarism:copyright of trained data ) These systems are often trained on large data sets, which can contain biases and other issues that can affect the output of the system. Without transparency in the data sets, it can be difficult to identify and address these issues.

Another issue is the lack of transparency in the algorithms used in generative AI systems. (Re: Tim’s lecture) A third issue is the lack of transparency in the decision-making processes used by generative AI systems. These systems can make decisions based on complex calculations and analysis, but it can be difficult to understand how these decisions are being made. This can be particularly concerning in cases where the system is making decisions that can have a significant impact on people's lives, such as in medical diagnosis or hiring decisions.

Lack of transparency on the effects on humans/society(有人用生成的画面语音内容进行伪造仿造、剑桥公司操纵大选)

Reflections for transparency in social media and big data research, by Walsh (2019):

1. “The first recommendation is that we may need to take into account not just the impact on the individual under study but the broader impact any experiment might have on society...”（对社会的影响）

2. “The second recommendation is that ethics approval may be needed...”（在伦理上是否被允许）

3. “The third recommendation is that subjects of any experiment may need to be informed directly after the study about the results and their participation...”（实验对象是否被告知）

大数据研究对隐私、自主权、幸福的影响

Big data research on social media invokes many concerns – privacy (can the user opt out of the ‘researchers gaze’); autonomy (does the research make the users do things they won’t otherwise?); wellbeing (does the research have the potential to change mood/health outcomes?)... are these clear to the users?

图形用户界面, 应用程序

描述已自动生成OECD：principle of AI ethics

Soft law code: weak response or informing general law?

Controlling the flow of data:

\Privacy act

What does the Privacy Act Protect?

The Privacy Act regulates the way an individual’s personal information is handled.  
‘Personal information’ is any ‘information or an opinion about an identified individual, or an individual who is reasonably identifiable:   
OAIC considers that ‘personal information’ includes ‘voice print and facial recognition biometrics (because they collect characteristics that make an individual’s voice or face unique)’

How do firms justify collecting personal information of personal information?

Australian Privacy Principles – an APP entity

(may only solicit and collect personal information that is reasonably necessary for one or more of its functions or activities)

(may only solicit and collect sensitive information if the individual consents to the sensitive information being collected, unless an exception applies

(collects personal information about an individual must take reasonable steps to provide notice.)

GDPR: individual rights

图形用户界面, 应用程序

描述已自动生成

AI risk based categories

文本

描述已自动生成

EU AI Act :obligations

文本

描述已自动生成Some requirements on organizations using High-risk AI systems

Reading: Algorithmic Fairness Is Not Only a Technical Challenge

Since decisions on fairness measure and the related techniques for fair algorithms essentially involve choices between competing values, fairness in algorithmic fairness should be conceptualized first and foremost as a political question and be resolved politically. I suggest that one promising way forward is through democratic communication.

人们对公平的定义是不一样的：一家公司认为模型是公平的，忽略种族差异，另一家认为由于黑人白人的准确率差异，模型是不公平的。the idea of fairness in algorithmic fairness is far from being uncontested. The idea of Bfairness^ in algorithmic fairness is in many ways contestable, which present an immediate problem to achieving algorithmic fairness. Firstly, there is a growing number of definitions for what Bfairness^ in algorithmic fairness amounts to, and it seems unlikely for researchers and developers to settle on the definition of fairness anytime soon.

In this paper, I attempt to show that there are more to creating algorithmic fairness than a set of technical tasks, i.e., there is an important political dimension in the problem of algorithmic fairness due to the contentious nature of the ideas of fairness and the fact that a decision on fairness measure and the balance between fairness and performance is in effect about competing values. （we should use politic method AFR to resolve problem of fairness）

Accessibility

Basically, technology is accessible if it can be used as effectively by people with disabilities as by those without(技术对于残疾人和非残疾人都是accessible的)

Accessibility refers to the degree to which an interactive product is accessible by as many people as possible. A focus is on people with disabilities

Universal Usability = a “design for all”

Accessibility goes beyond just ’catering for those with disabilities’.

• Situational impairments

• Consider: a busy parent during the breakfast rush

• Consider: defense personnel during deployments in a humanitarian crisis

• Consider: remote learning/work during the Covid-19 pandemic

• Temporary disability/temporary impairment

• Consider: a student who broke their arm after a bicycle accident

• Consider: a lecturer who has a spinal injury

Equity

the quality of being fair and impartial: equity of treatment.

Create an algorithm to divide a finite pool of resources (X) equitably across N participants (P1, ... PN).

Example answer: Equal Share Algorithm

• Calculate share = (X / N)

• For each person in participant pool {P 1 , ... P N }:

• Allocate current person their equal allocation (share)

Now consider that the algorithm is to be deployed in the real world to automate the allocation of resources to different communities.

For a given affluent community, assume everyone is sufficiently well-off and have more than enough resources, money etc EXCEPT for two people (only P1 and P2 ).

P1 and P2 are the only ones who needs access to resources (food, water, etc) due to (hunger, health conditions, etc)

Suddenly, your equitable algorithm doesn’t seem so equitable after all.

The design of an automated / computerised / AI-driven system can seem fair...but: Yet, these algos might violate equity (and accessibility) AFTER they are deployed

We only notice the problem when we deploy it... and only then find out that it doesn’t work in certain cases

Systems are inherently complex: what works in isolation does not work ‘as a whole’, or even when deployed in circumstances (external factors, e.g., social factors) we did not foresee.

With machine learning, we need vast amounts of complex data when building the systems as well. Feedback loops + complexity = bad

The problems ‘after deployment’ get fed back into the system to entrench these issues.(模型部署后的反馈会进一步加深这个问题，outputs of today become the inputs(of training model) tomorrow )，Again, the complexity of modern systems make these hard to untangle。

写字比赛软件：

Alice手断了不能写，好几个月后排名一直下跌（accessibility issue）

Elijah是左撇子，写字很漂亮，但是系统是利用右撇子的作品训练出来的，不符合右手特性，所以他的分数低30%（equity issue）

Gender bias in hiring algorithms can occur in 3 forms:

1. bias in datasets(lack of diversity from internet: diminish cultural perspectives and entrench bias within language models) 2.bias in the system 3.bias in human decisions

生成式AI的歧视

大多数生成的图片：医生就是男的，护士就是女的，图书馆管理员就是女的

新版本在性别方面的偏向会好一点，但是宗教、族裔、年龄上的变化仍然很少，都是年轻的，黄种人

Equality：gender bias for hiring system of Amazon

Amzaon 案例: Human shortlisting of candidates – reflects human/societal biases.Training a classifier à model that entrenches the bias

Even though the algorithm can be opened up for auditing, and e.g. just uses

established, off-the-shelf packages/techniques – the ENTIRE SYSTEM needs to be interrogated（不仅算法需要审计，整个系统都需要审计，outputs of today become the inputs(of training model) tomorrow）

also: Also, don’t forget **accessibility:** If the system was deployed and everyone had to apply using a web-based system, say,...What about accessibility issues for people with disabilities, situational impairments, etc.?

the idea of the Original Position (OP), proposed by political philosopher John Rawls [21]. The “most appropriate moral conception of justice” [11] is obtained when the parties take up the “veil of ignorance(无知的面纱)”, completely depriving themselves of all knowledge of their own personal circumstances and attributes（完全剥夺他们对自己个人情况和属性的所有知识）; in short, putting themselves in the shoes of others（设身处地为他人着想）

不是所有的fairness/equity问题都能归咎于算法和数据集

重要的是repair——产品坏了就丢，这是过早放弃：没钱的人没法换新。消费者文化也会导致这个现象（算法可能会导致不公平——算法的更新迭代使得旧版本的手机无法使用，不得不购买新手机）

Reading: implication of Amazon: Our argument is that any amount of initial bias – no matter how little -- will perpetuate in the algorithmic models used, further continuing the bias, and thus reinforces it by keeping human oversight out of the loop.

When ranking algorithms for recruitment are trained with biased data sets, the technology impacts the organisation in a way that reflects the organisational operation, while at the same time influencing the way it operates. This means hiring algorithms trained with biased data can replicate existing inequalities while *also* introducing new ones.

Solution on design of AI:

1. ***Prioritising explainability and interpretability over commercial accuracy***

2. ***User-centred design and domain expert knowledge :***Our data scientist is required to observe, ask questions and solicit input from panel participants – including assumptions, justifications, and clarifications – in order to gain domain knowledge of the task at hand. This meant the hiring panel rather than the data scientist drove the logic of the algorithm specifications, while the data scientist’s role is on the technical facets - such as techniques for data wrangling, choice of model, and programming decisions

3. ***Use of open source tools as opposed to black-boxing***

Solution gender Bias in AI:

1. Provide training programs to introduce human resource professionals to the potential of gender bias in algorithmic judgements in hiring process.
2. Complete regular audits of hiring by gender across all positions to identify potential roles that are vulnerable to gender discrimination.
3. Create established quota systems for hiring to ensure women are not excluded from male-dominated or gender balanced professions based on hiring biases.
4. Create proprietary hiring algorithms that are transparent and trained with the aim of reducing gender bias in hiring, with regular audits of algorithm output and models (both trained on human judgment as well as unsupervised).

Data governance

Data governance represents the program used by [an organisation] to manage the organisational bodies, policies, principles, and quality that will ensure access to accurate and risk-free data and information..

The Data Governance Institute defines data governance as "a system of decision rights and accountabilities for information-related processes（信息相关流程的决策权和责任系统，根据商定的模型执行）, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods.（谁可以对数据采取什么行动，合适、何种情况使用什么方法）

对于数据使用和数据来源的监管、计划、控制。

Key responsibilities of data governance:

文本, 信件

描述已自动生成图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

用户和大型tech company的关系是不对称的，他们拥有大量的用户数据，

在FB、CA、Meta掌握大量数据的情况下

在kant角度下，如果你不做什么，每个人也都什么都不做，他们会持续使用隐私数据，持续干涉大选。所以do nothing is bad,所以阻止这些公司，保护隐私是你们所有人的duty

GDPR data protection

文本, 信件

描述已自动生成

“data protecnon is something you now have to consider whenever you do anything with other people's

personal data”, not just an aqerthought

• “People have the right to see what personal data you have about them and how you're using it” Though in pracnce, this is tricky...

• “You are also required to quickly communicate data breaches to your data subjects” and not just keep quiet about it!

• NB: A ‘right’ to explainability – this is the subject of many scholarly works in legal studies and digital ethics. (Our very own Prof Tim Miller will be able to provide the technical details about what it means to have explainable AI)

lack of transparency in the data sets used to train generative AI systems”

module9的reading要看

data论文：

data collection\annotation documentation practice

dataset是虚假的存在bias的（在训练集上用了大量来自西方世界的数据，在模型训练中缺乏解释性，rely on cheap tricks’2 rather than human-like reasoning capabilities）

In closing, we advocate for a turn in the culture towards carefully collected datasets, rooted in their original contexts, distributed only in ways that respect the intellectual property and privacy rights of data creators and data subjects, and constructed in conversation with the relevant scientific and scholarly fields required to create datasets that faithfully model tasks and tasks which target relevant and realistic capabilities.

Facebook关闭的影响：These stakeholders are: *dependent communities,* in particular the socio-economic and media ecosystems that depend on Facebook to flourish; *existing users*, (active and passive) individuals, as well as groups, whose data are collected, analysed and monetised by Facebook, and stored on the company’s servers; *non-users*, particularly deceased users whose data continues to be stored and used by Facebook, and who will represent hundreds of millions of Facebook profiles in only a few decades; and *future generations,* who may have a scientific interest in the Facebook archive as a historical resource and cultural heritage.

Recommendation:

DEVELOP A REGULATORY FRAMEWORK FOR SYSTEMICALLY IMPORTANT TECHNOLOGICAL INSTITUTIONS.

STRENGTHEN THE LEGAL MECHANISMS FOR USERS TO CONTROL THEIR OWN DATA IN CASES OF PLATFORM INSOLVENCY OR CLOSURE.

**STRENGTHEN LEGAL PROTECTION FOR THE DATA AND PRIVACY OF DECEASED USERS.**（加强立法保护对于去世用户的数据）

CREATE STRONGER INCENTIVES FOR FACEBOOK TO SHARE INSIGHTS AND PRESERVE HISTORICALLY SIGNIFICANT DATA FOR FUTURE GENERATIONS（鼓励facebook分享观点并且为未来一代保存历史数据）

Explainability

Explainable AI =understanding AI models and decisions

Explanation answers a :why question

Who cares explainable AI? When? why ?

Data scientist: Understand the model. Debug the model. Improve model performance

Business Owner: Understand the model. Evaluate model suitability. Accept model use

Risk Modeller : Challenge the model. Ensure model robustness. Approve model.

Regulator: Check impact of model on consumers. Verify model reliability.

Consumer :“What is the impact on me?” “What actions can I take?”

The challenges of explainable AI.

(human always focus on causality, but machine focus on correlation.)

(the human problem “complexity VS simplicity” trade-off)

可解释AI方法的特性：（intrinsic VS post-HOC）（内在的VS事后的）

图示

描述已自动生成

图示

描述已自动生成

Model specific: inner workings of model used for XAI

Model AGNOSTIC(与模型无关的): use only inputs and outputs for XAI. Applicable to any model with that interface.

Explainability：

ATTRIBUTION-BASED EXPLANATIONS

EXAMPLE-BASED EXPLANATION: PROTOYPES

RULED-BASED EXPLANATION（EXTRACT RULES POST-HOC LEARN INTERPRETABLE RULES DIRECTLY）

CONTRASTIVE EXPLANATION

Reading：(XAI :One step to improving people’s trust in these algorithms, and ultimately, to produce ethical artificial intelligence, is to produce artificial intelligence that can explain why it made a decision.

XAI **challenges:1.opaqueness:** Deep neural networks, are highly opaque. Not only will the engineer have a hard time predicting the output, but they will have a hard time tracing the reasoning back and debugging.

**Solution: a.Importance weighting.** Many techniques reverse engineer which parts of an input were “important” for a decision. b. **Interpretable models.** Other techniques aim to extract or learn a less opaque (more interpretable) model; c. **Discard deep neural network models.**

**Challenge 2: Causality.**

**Solution: 1.Learn causal models.2. Exploiting human strengths** 3. **Human-centeredness.**

AI, Art and Digital Ethics

Replication:

We say that a generated image has replicated content if it contains an object (either in the foreground of background) that appears identically in a training image, neglecting minor variations in appearance that could result from data augmentation"

language models have the ability to memorise their data  
• instances of images being reproduced

Bias and stereotypes: "The AI convo is changing - "AI" now being used colloquially to refer to an inferior copy, not a supercharged

version of the original"(AI版本已经被认为是低等的拷贝而不是原版的加强)

Art + Technology: the possibilities

Art is a lie that makes us realize truth.

Something that is created with imagination and skill and that is beautiful or that expresses important ideas or feelings.

图形用户界面, 文本, 应用程序

描述已自动生成Art is not just imaged and pictures:

Not the end of Art: Art is an adventure into an unknown world, which can be explored only by those willing to take risks.(仔细看有明显的错误，扭曲以及不协调的部分)

AI art will make designers obsolete

Ubuntu philosophy and digital ethics

Inclusion in digital ethics

Anglo-Saxon traditions moral theories: emphasise people’s individual rights and responsibilities.

Digital governance mainly focuses on protecting autonomy within a private sphere

African, Asian and Latin American societies have values and cultural traditions that can be characterized as less individualist

Low-income countries tend to be more collectivistic.(低收入国家更加注重群体)

Individualistic societies emerge with economic growth (个人社会主义随着经济增长而出现)

Rational beings are called persons inasmuch as nature already marks them out as ends in themselves(kant)

Community or individual always involves dynamically and mutually constituting other humans.

Collectivistic:

African relational humanism:a person is a person because of other people.

Whatever happens to the individual happens to the whole group, and whatever happens to the whole group happens to the individual. The individual can only say: ‘I am, because we are and since we are, therefore I am.

An individual must subject themselves to their community to qualify for personhood(一个人必须服从他的社区才能获得人格)

Individualist logic can assume: communalism is incompatible with evolving societies.

Group thinking is incompatible with individual liberties

Collectivism cannot be egalitarian or ensure that all people have equal rights and opportunities.

African personhood features individual agency within communal obligation: a relational and positive sense of autonomy, which involves the community helping or guiding one to use one’s ability and knowledge of one’s social relations and circumstance to choose freely the requisite goods for achieving one’s life plan.(社区帮助指导个人，利用自己的能力和社会关系和环境来自由选择自己的生活计划所必需的物品)

African personhood :

should not be considered a fixed universal concept

Is contextual and creatively constituted.

Africans welcomed Europeans into their communities from early on.

Hip-hop artists: produce music in the backyard of low-income townships.( Differentiate themselves from white people or middle-class black people)

绿色的标志

描述已自动生成

Relational sense of self-worth situated in a person’s relationship with others(个人与他人关系中的自我价值感)：主要关注于给予年轻人建议，帮助社区中的人成长已经成为一种义务。

通过在社交网络的活跃互动来获得声望，which aligns with a person’s connectedness to their community and their respect for others

电器维修案例：三个店铺，在与他人的关系中寻找自我价值，分享工具、客户、电力、知识和解决方案。通过当地的社会关系来管理不确定性

Interconnections between local, national and international systems enhance productivity and manage precarity。

Traditional community: a spatially compact set of people with a high frequency of f2f interaction and many interconnections

Contemporary community: a set of people who meet f2f intermittently or interact entirely via phones/the internet with

Difference between concrete and imagined sets of people: psychological sense in which people identify with others and have close and sympathetic social relations they perceive to be similar(人们认同他人并拥有他们认为相似的密切和同情的社会关系的心理意义)

A shared physical and social context supports values, norms, experiences, places, identity, collective goals and community members play multiple roles by participating in different activities.

Maimaila Community Network , South Africa

Ancestry and everyday life contribute shared values, norms, experiences, collective goals and people play multiple roles in the community(祖先、每天的生活贡献着共同的价值，准则和经验还有共同的目标)

Community network:

Network: tech infrastructure e.g., solar-powered Wi-Fi with internet connections

Community: people who set-up ,manage, use and decide about the network.

家庭合作去支持失去亲人的家庭，为葬礼提供人力、资金和装备支持。

设计一个App去支持Leola来展示集体工作：

资金的收集和分配、长者建议的共享、识别影响社区不同方面的挑战，并且分配给不同的人去解决。与当局联络并形成社交网络（liaise with authorities ）

MCN利用Matshepo’s技术：解放黑人内心，治愈殖民与种族隔离思想造成的伤害。

如果不联手，并且没有这样的环境，转型是不会发生的。

Ubuntu applied in different disciplines:

Moral tension between civil liberties and public health arising from the state’s decision-making during pandemics(疫情期间国家决策引发的公民自由和公共卫生之间的道德紧张)

Orthopaedic practices enabling respectful and beneficial care for patients without reducing the common good(骨科原则 使得对病人尊重和有益的照顾没有减少共同利益)

戴口罩，保护自己的同时保护别人

在技术中：人格先于人际关系存在的个人主义逻辑占主导。

To promote shared humanity, ethics needs to go beyond technical solutions and account for existing privilege and future empowerment

Ubuntu emphasizes that humanity emerges in relations between people. Thus, when people in one social group violate the humanity of people outside of that group, they deprive their own group of humanness(Ubuntu 强调人性出现在人与人之间的关系中。 因此，当一个社会群体中的人侵犯了该群体之外的人的人性时，他们就剥夺了自己群体的人性)

To be human ourselves we must prioritise other people’s dignity and well-being especially for the weakest(若要成为人就先要考虑他人的尊严和福祉，尤其是对于最弱的人)

Algorithmic and data injustices:

AI 模型复制并且放大了历史和社会经济的不平等：显性和隐性的刻板印象（种族、性别等）———伤害了他人的尊严和福祉，破坏了Ubuntu的规则。（在训练数据中存在bias的原因：优势群体收集了大量有利于他们自己的数据进行训练，忽视了那些弱势群体的经验）

算法设计者和采集数据的社区或者算法被实施的社区的联系不够紧密，是断开的。

Ubuntu ethic frameworks promote shared humanity: (该框架是如何促进人性共享的)

AI系统的伦理框架：撤销之前和现在的社会不公平，评估弱势群体，帮助实施促进人文主义价值观的价值。（Assist in implementing values that promote values of relational humanism）

Values of African relational humanism(非洲关系人文主义的价值观)

The Common good:(Emphasise communal aspects of life in understanding existential situations; Do not undermine communal relationships or social cohesion; Consider the well-being of the community over the individual; People who identify with a communal relation know that others will act in their own best interest via the interest of their community)

The common humanity of humans despite their diversity(富有多样性的共同人文价值观)

Prioritizing the dignity of others, particularly the most vulnerable（优先他人的尊严，尤其是弱势群体）：(Collect diverse data from different regions, races, ethnicities, tribes, geographies and socioeconomic groups；Use data in ways that are sensitive to diverse needs and experiences;

Consider Ubuntu in managing conflicting cultural values：

（1.Focuses on a person’s free and voluntary decision,based on sufficient knowledge and without coercion自愿无强迫;2.基于kant对人格的概念：人的决策的独立自主权3.提出的问题是基于AI能力的问题：比如算法复杂性、抓取互联网数据的合法性）)

管理价值观冲突的例子：知情同意：ubuntu的概念首先是尊重人的内在尊严无论他们的判断和他们的社区是怎样的。需要相互关系中的知情同意：如果一个人对社区中是团结一致的，他们在互动中是透明的并且也是优异保护社区成员内部各自的利益。）

Communal relations: generosity, hospitality, compassion, friendliness and sympathy for others（共同的关系：慷慨、热情、同情、共情）

Harmony and peaceful relations

Tolerance and mutual respect of humans despite their diversity. Solidarity, reciprocity and cooperative participation. Resolving difficult moral decisions through consensus

Trust

Ubuntu embeds a moral sense of social trust; that is, confidence in another person’s actions through long-term, honest, kind and respectful relationships(Ubuntu 凸显了社会信任的道德观念，对他人行动的信任，通过长期诚信尊重友善的关系)

映射到AI系统中：Achieving this with AI requires:

1.Long-term collaboration, involving communities and experts in designing, developing and overseeing AI systems to gain sufficient understanding to be able to judge them（社区和专家在设计、发展、监控AI系统方面的长期合作，去获得充分的理解并去评估AI系统的正确性）

Transparent communication about the capabilities and limitations of AI tools e.g., reporting cases where an algorithm fails（对于AI工具的能力和缺陷建立透明畅通的沟通机制）

Monitoring, reporting on and addressing negative consequences（监控报告消极后果）