Mitigating Algorithmic Bias

Or, Responsible Data for London

Hello!

Rachel Coldicutt

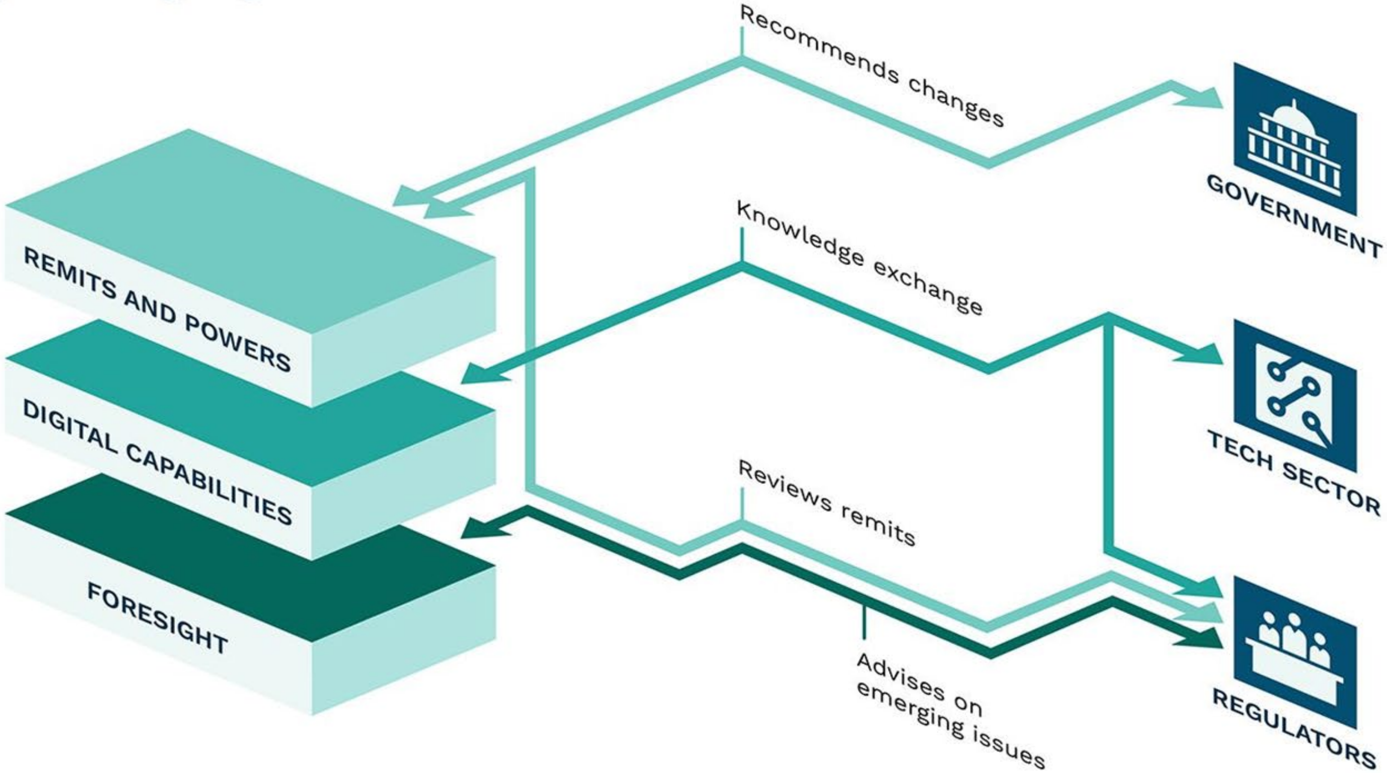
CEO, Doteveryone

Discussion

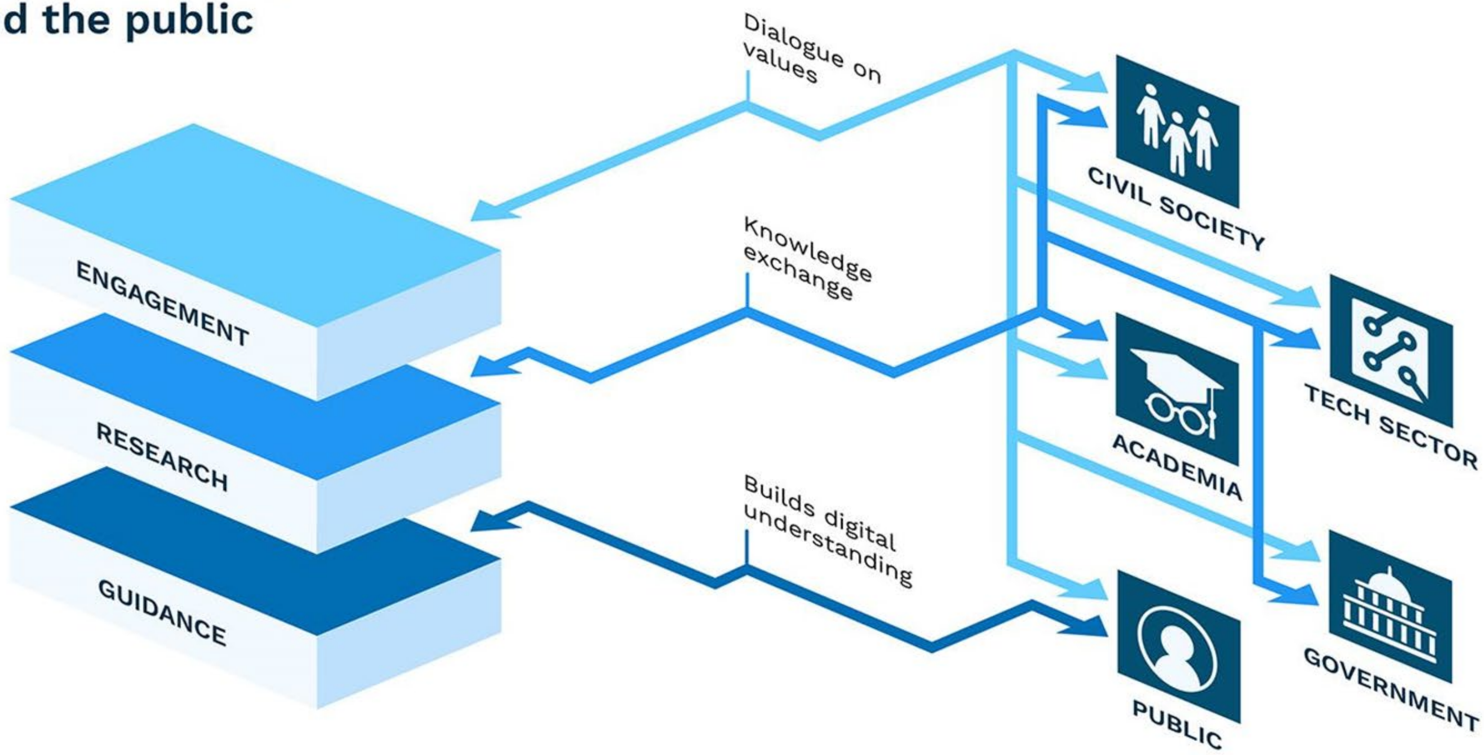
1. What work is still needed to develop proposals for strengthening digital regulation?

1. Who should be responsible for making these proposals a reality?

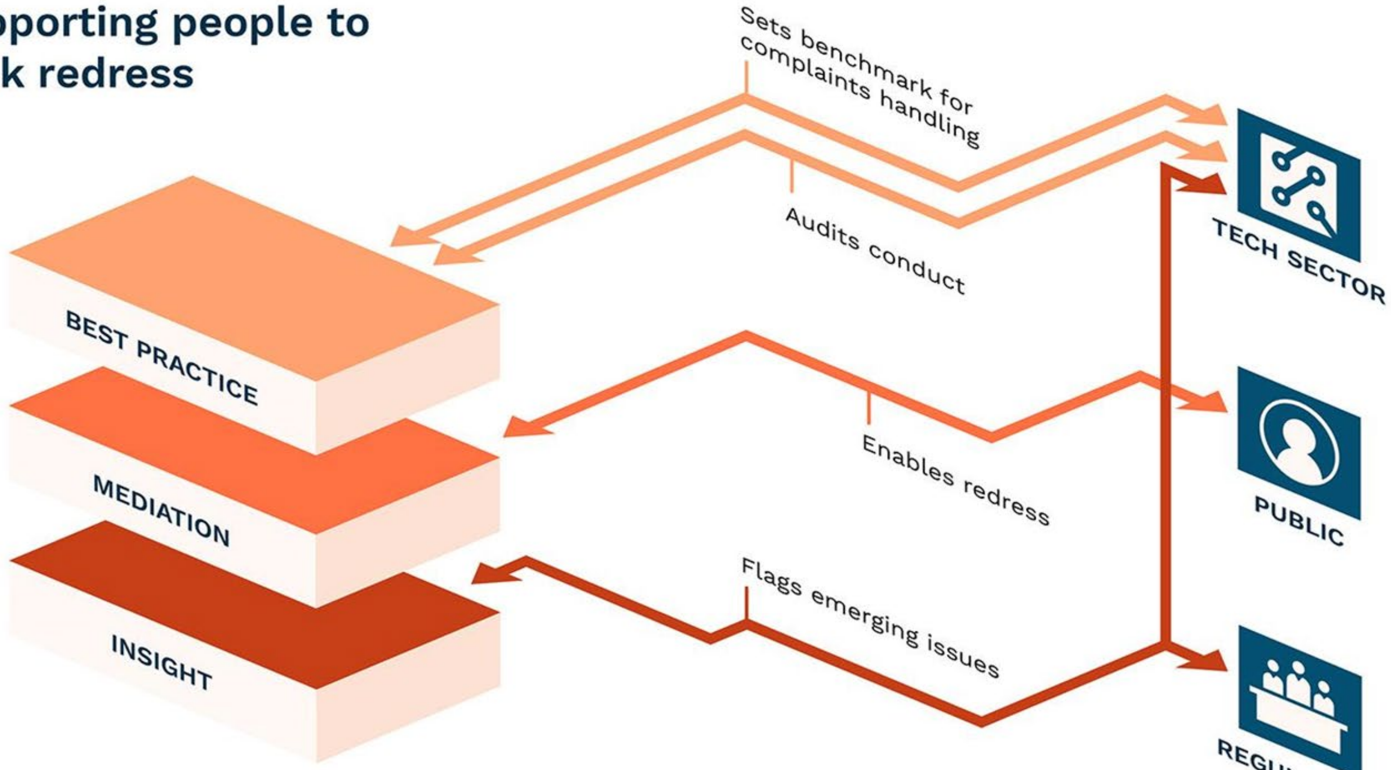
Empowering regulators



Informing policymakers and the public



Supporting people to seek redress



Doteveryone is calling for a new Office for Responsible Technology to:

1. **Empower regulators.** The Office sits above existing regulators, identifies the gaps in regulation and supports regulators with the expertise to respond to digital technologies as they affect their sectors.

1. **Inform the public and policymakers.** The Office creates an authoritative body of evidence about the benefits and harms of technologies to underpin the work of regulators, builds public awareness, and engages all parts of society to create consensus around a future vision for technology to underpin the regulatory system.

1. **Support people to find redress.** The Office ensures the public can hold owners of technologies to account for individual and collective harms derived from their use, setting best practice in online harm prevention and enabling backstop mediation when these standards are not met.

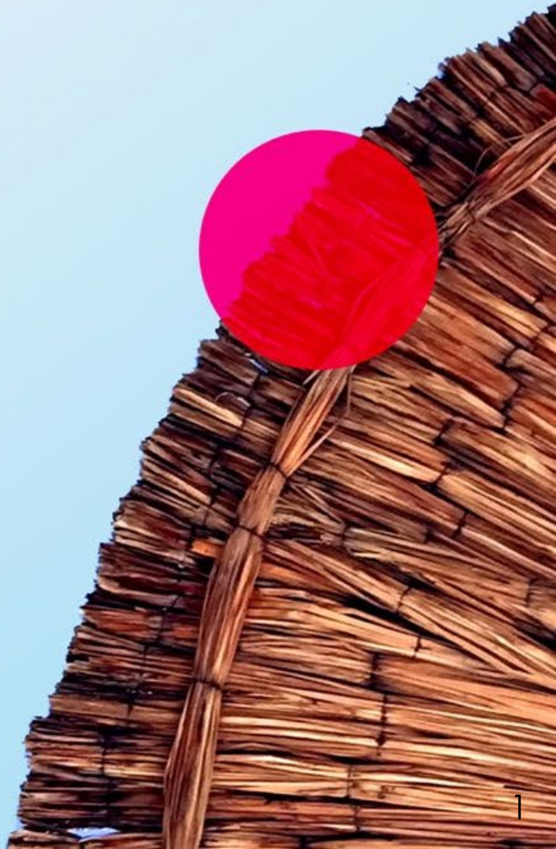
Discussion

1. What work is still needed to develop proposals for strengthening digital regulation?

1. Who should be responsible for making these proposals a reality?

A brief introduction to public sector AI ethics and its challenges

Eddie Copeland
Director of Government Innovation



Artificial Intelligence (AI) is the science of using machines “to do things that traditionally required the human mind”

nesta

Why's it worth the hassle to get right?

AI has the potential to:

- Codify best practice and roll it out at scale
- Remove human bias
- Enable evidence-based decision making in the field
- Spot patterns that humans can't see
- Optimise systems too complex for humans to model
- Quickly digest and interpret vast quantities of data
- Automate cognitive activities that require significant human effort

Specific existing use cases

AI is already being used by governments and public sector organisations for specific activities such as:

1. Analysing case notes to determine whether a child is likely to be taken into care
2. Spotting tumours in X ray scans
3. Identifying risky behaviour from CCTV footage
4. Predicting where certain crimes are likely to occur
5. Detecting fraudulent benefits / tax claims
6. Optimising traffic intersections
7. Enabling smart chatbots to answer citizen questions

One of the most important - and immediate - areas of use for AI in the public sector will be in enabling **algorithmic decision making**: decisions made or informed by machines.

Ethical concerns

Critics have raised ethical concerns that AI could be used by governments and the public sector in ways that **invade privacy; or cause harm, unfairness and moral wrongs**.

Calls have been made for new codes, standards and principles to be created.

Examples of concerns raised about the use of AI

These ethical concerns - and many proposed solutions to them - tend to focus on one or more of the following **three stages** of deploying an AI.

CREATION	FUNCTION	OUTCOME
How the AI is created	**How the AI works**	**What the AI is used to do**

Examples of ethical concerns raised about the use of AI

CREATION	FUNCTION	OUTCOME
Does the AI use data that invades individuals' right to privacy?	Are the assumptions used by the AI correct?	Is the AI being used to do something unethical?
Is the training data accurate and truly representative?	Are the factors used by the AI to make a decision reasonable and fair?	Is anyone responsible / accountable if a negative outcome is produced by an AI?
Does the training data contain historic biases that could be perpetuated?	Can anyone see and understand how the AI works and audit how a given output was created?	Will people know if a decision affecting them was made by an algorithm?
What happens when the use of the AI renders the training data out of date?	Can we be sure the AI is protected against hacking and manipulation?	What recourse will people have if an AI discriminates against them or causes them harm?

It's important to acknowledge that many of these concerns have a strong rationale. For example, concerning the creation of algorithms using AI:

- The public have shown they are concerned by the way their data - and especially their personal data - is being used by AI, as seen in corporate examples such as the Facebook / Cambridge Analytica scandal.

- Some public sector applications of AI have been found to discriminate based on biases in the training data - e.g. US prisons' use of past parole data.

- Some public sector uses of AI have been criticised for using data about factors such as race or religion, which many feel is misleading, inappropriate, or unethical.

Concerning how an AI functions, there are legitimate concerns about algorithms' operation and opacity:

- The assumptions on which an algorithm are based may be broadly correct, but in areas of any complexity they will at best be incomplete.

- The code of algorithms may be unviewable in systems that are proprietary or outsourced. This is known as the '**Black Box**' problem.

- If the code is viewable and comprehensible, some worry that this will make it easier for malicious hackers to manipulate the algorithm.

Lastly, there are concerns about the ways in which algorithms might be used:

- Algorithms have been used in inappropriate contexts, such as companies using job applicants' credit scores to determine whether to hire them.

- Algorithms may be deployed without appropriate human oversight leading to actions that could cause harm and which lack accountability.

- The scalability of algorithms means that any negative impacts could be far reaching (see Cathy O'Neil)

Public sector as a special case?

- Monopoly provider of the services it offers
- Interacts with very vulnerable people
- Decisions may have significant consequences on a person’s life
- Democratically elected governments have special duties of accountability

nesta

Examples of codes, standards and principles for the ethical use of AI

Google AI Principles

1. Be socially beneficial
2. Avoid creating or reinforcing unfair bias
3. Be built and tested for safety
4. Be accountable to people
5. Incorporate privacy design principles
6. Uphold high standards of scientific excellence
7. Be made available for uses that accord with these principles

See: https://www.blog.google/technology/ai/ai-principles

Microsoft AI principles

1. **Fairness - AI systems should treat all people fairly**
2. **Inclusiveness - AI systems should empower everyone and engage people**
3. **Reliability & Safety - AI systems should perform reliably and safely**
4. **Transparency - AI systems should be understandable**
5. **Privacy & Security - AI systems should be secure and respect privacy**
6. **Accountability - AI systems should have algorithmic accountability**

See: https://www.microsoft.com/en-us/ai/our-approach-to-ai

UK Government initial code of conduct for data-driven health and care technology

Department of Health & Social Care

1. Define the user
2. Define the value proposition
3. Be fair, transparent and accountable about what data you are using
4. Use data that is proportionate to the identified user need (data minimisation principle of GDPR)
5. Make use of open standards
6. Be transparent to the limitations of the data used and algorithms deployed
7. Make security integral to the design
8. Define the commercial strategy
9. Show evidence of effectiveness for the intended use
10. Show what type of algorithm you are building, the evidence base for choosing that algorithm, how you plan to monitor its performance on an ongoing basis and how you are validating performance of the algorithm.

See: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology

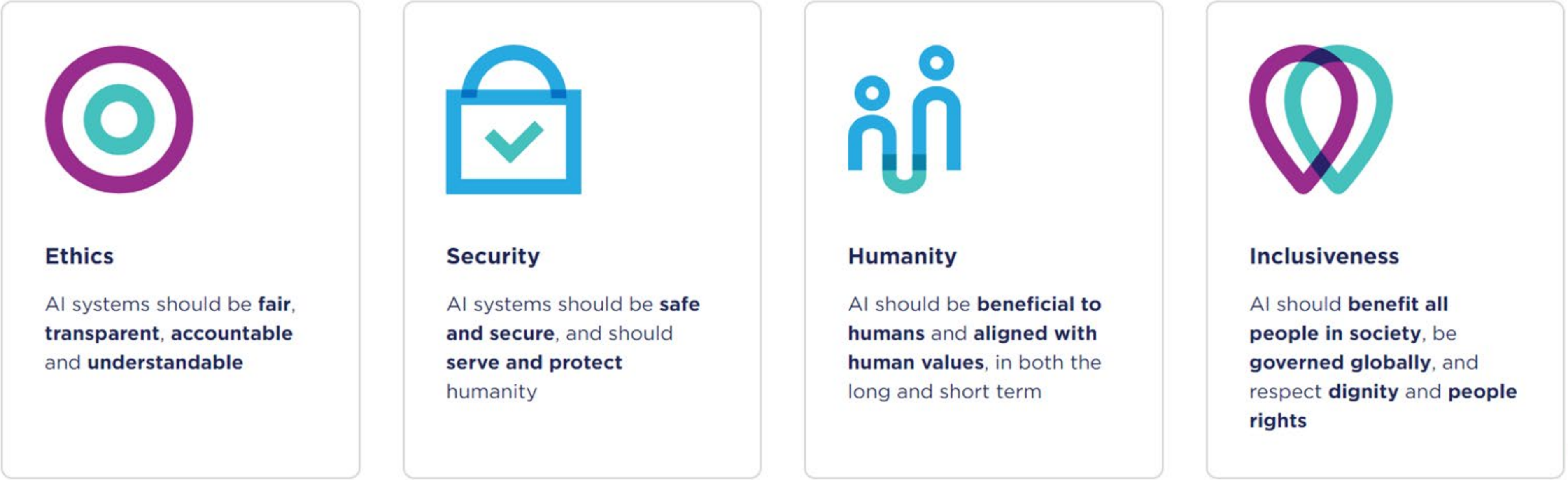
European Commission’s High Level Expert Group on Artificial Intelligence

“Trustworthy AI” founded on fundamental principles of individuals’ rights

1. Accountability
2. Data Governance
3. Design for all (by all - include diversity)
4. Governance of AI Autonomy (Human oversight)
5. NonDiscrimination
6. Respect for Human Autonomy
7. Respect for Privacy
8. Robustness
9. Safety
10. Transparency

See: https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

Smart Dubai - AI Principles



Ethics

AI systems should be **fair**, **transparent**, **accountable** and **understandable**

Security

AI systems should be **safe and secure**, and should **serve and protect** humanity

Humanity

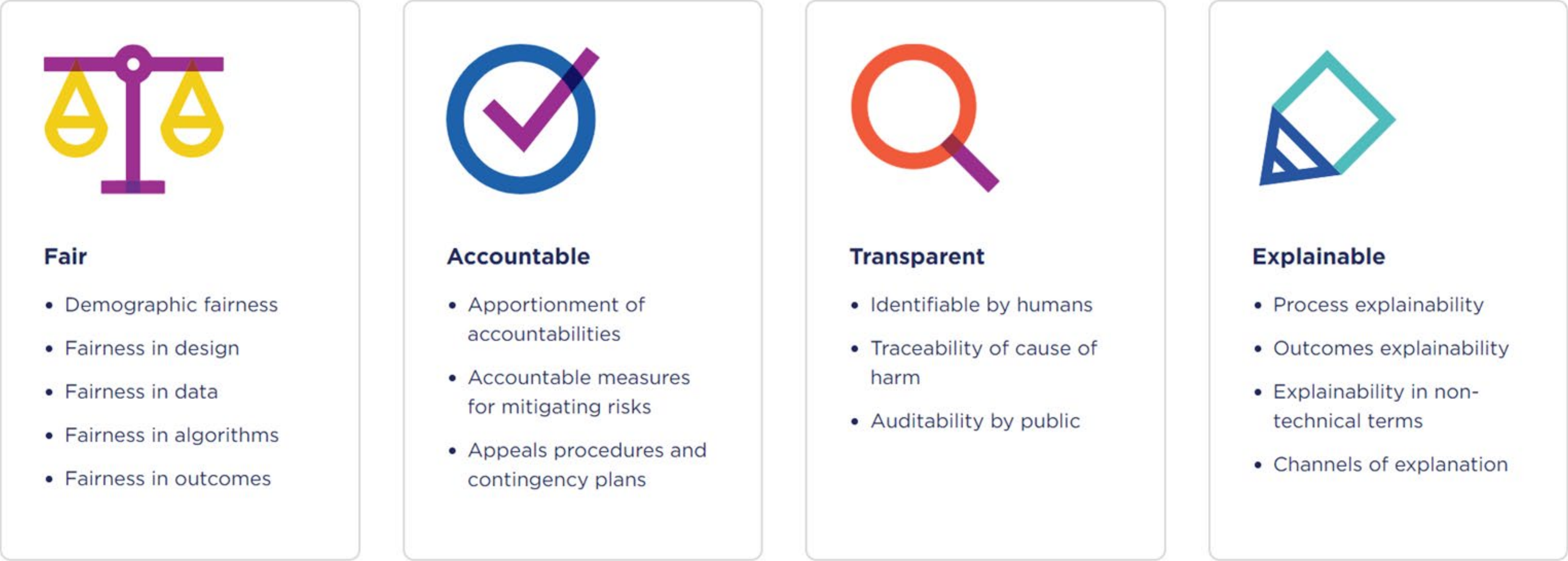
AI should be **beneficial to humans** and **aligned with human values**, in both the long and short term

Inclusiveness

AI should **benefit all people in society**, be **governed globally**, and respect **dignity** and **people rights**

See: https://smartdubai.ae/initiatives/ai-principles-ethics

Smart Dubai - AI Ethics



Fair

- Demographic fairness
- Fairness in design
- Fairness in data
- Fairness in algorithms
- Fairness in outcomes

Accountable

- Apportionment of accountabilities
- Accountable measures for mitigating risks
- Appeals procedures and contingency plans

Transparent

- Identifiable by humans
- Traceability of cause of harm
- Auditability by public

Explainable

- Process explainability
- Outcomes explainability
- Explainability in non-technical terms
- Channels of explanation

See: https://smartdubai.ae/initiatives/ai-principles-ethics

Nesta principles for public sector use of AI

1 - Every algorithm should be accompanied with a description of its function, objectives and intended impact, made available to those who use it.

2 - A description of the data on which an algorithm was trained and the assumptions used in its creation should be published, together with a risk assessment for mitigating potential biases.

3 - A list of all the inputs used by an algorithm to make a decision should be published.

4 - Citizens must be informed when their treatment has been informed wholly or in part by an algorithm.

5 - Every algorithm should have an identical sandbox version for auditors to test the impact of different input conditions.

6 - When using third parties to create or run algorithms on their behalf, public sector organisations should only procure from organisations able to meet Principles 1-5.

7 - A named member of senior staff (or their job role) should be held formally responsible for any actions taken as a result of an algorithmic decision.

8 - Public sector organisations should commit to evaluating the impact of the algorithms they use in decision making, and publishing the results.

What do we make of these different codes, standards and principles?

There's lots of overlap in the recommendations from codes, standards and principles:

CREATION	FUNCTION	OUTCOME
Reveal / publish the training data	Make the code of the AI transparent and open for inspection	Ensure intended and actual outcomes are fair, transparent and aligned with human values
Identify and minimise bias in the training data	Do not create or procure black box AIs	Ensure outcomes can be explained
Respect privacy / don't use data in ways that are creepy	Identify and minimise bias and limitations in the AI's assumptions	Ensure there is a process of oversight and evaluation
Do not use data on sensitive factors such as race and religion	Ensure the factors and function of the AI can be explained	Ensure a person is accountable for decisions made using the AI
Use personal data in compliance with GDPR	Offer identical sandbox versions of the AI to enable testing	Ensure outcomes are fair, inclusive and respect dignity and rights
	Ensure fairness by design	Ensure use of AI is known and there is a process of appeal
	Protect from manipulation and hacking by design	Mitigate against harms

But it’s not that simple…

nesta

Different levels of complexity of AI have different consequences for ethics

Complexity

3	Unlimited quantities of unstructured training data such as video, photos, sound, free text	Dynamic model constantly evolving based on live data	Model used as one part of long and complex decision-producing chain
2	Defined quantity of structured or unstructured training data used for a one-time creation of model	Static model created using one-time machine learning process	Model used to cover simple and clearly defined point of decision making process
1	Defined number of structured datasets used for one-time weighting of model	Static model using human inputed rules weighted by machine learning	Model used to cover simple and clearly defined point of decision making process
	CREATION	FUNCTION	OUTCOME

0101
1001

nesta

Understanding level 1

1	CREATION	FUNCTION	OUTCOME
	Defined number of structured datasets used for one-time weighting of model	Static model using human inputed rules weighted by machine learning	Model used to cover simple and clearly defined point of decision making process

The simplest form of AI involves using a small number of structured datasets to correctly weight a number of factors that are deemed to be important by humans.

Example: Firefighters could be asked to detail the factors their experience has told them are relevant to a building’s risk of fire. Datasets can then be sought that relate to those factors in order to train an AI. Machine learning is used to weight the factors according to the extent they are predictive of a high risk building.

Understanding level 2

2	CREATION	FUNCTION	OUTCOME
	Defined quantity of structured or unstructured training data used for a one-time creation of model	Static model created using one-time machine learning process	Model used to cover simple and clearly defined point of decision making process

Instead of merely weighting factors deemed relevant by humans, a level 2 AI increases complexity by deciding for itself what factors are relevant, how they should be weighted, and how they lead to a given outcome.

Example: A local authority could use machine learning to analyse thousands of free text social worker case notes about vulnerable children to spot patterns and correlations that predict which of them are most likely to be taken into care in the future.

Understanding level 3

3	CREATION	FUNCTION	OUTCOME
	Unlimited quantities of unstructured training data such as video, photos, sound, free text	Dynamic model constantly evolving based on live data	Model used as one part of long and complex decision-producing chain

In the most advanced forms of AI, neither the training data nor the models created are static. The model is continuously updated based on new data, that will often be vast in a quantity and unstructured.

Example: A police surveillance system constantly analyses CCTV footage and sound from dozens of train stations in order to spot suspicious behaviour.

Challenges for ethical approaches

It it really possible to be transparent about the training data and assess for bias if it’s unlimited and unstructured - e.g. thousands of hours of CCTV footage?

And can we meaningfully talk about ‘explainability’ if not even the developers of an AI know how it reasons? See example of AlphaGo AI, which played Go against itself to learn the optimal strategy.

Challenges for ethical approaches to Levels 2 and 3

In the tables that follow, for each recommendation:

✓ indicates it's straightforward / possible

✗ indicates it's extremely hard / impossible

~ indicates it's only possible in some circumstances

Viability of ethical recommendations for the 3 levels of AI: Creation

CREATION	Level 1	Level 2	Level 3
Recommendations	Defined number of structured datasets used for one-time weighting of algorithm	Defined quantity of structured or unstructured training data used for a one-time creation of algorithm	Unlimited quantities of unstructured training data such as video, photos, sound, free text
Reveal / publish the training data	✓	✓	~
Identify and minimise bias in the training data	~	~	~
Respect privacy / don't use data in ways that are creepy	✓	✓	~
Do not use data on sensitive factors such as race and religion	✓	✗	✗
Use personal data in compliance with GDPR	✓	✓	~

nesta

Viability of ethical recommendations for the 3 levels of AI: Function

FUNCTION	Level 1	Level 2	Level 3
Recommendations	Static algorithm using human inputed rules weighted by machine learning	Static algorithm created using one-time machine learning process	Dynamic algorithms constantly evolving based on live data
Make the code of the AI transparent and open for inspection	✓	✓	✓ (but meaningless)
Do not create or procure black box AIs	✓	✓	✓
Identify and minimise bias and limitations in the AI's assumptions	✓	~	✗
Ensure the factors and function of the AI can be explained	✓	~	~
Offer identical sandbox versions of the AI to enable testing	✓	✓	✗
Ensure fairness by design	✓	~	~

nesta

Viability of ethical recommendations for the 3 levels of AI: Outcome

OUTCOME	Level 1	Level 2	Level 3
Recommendations	Algorithms used to cover simple and clearly defined point of decision making process	Algorithm used as one part of long and complex decision-producing chain	Algorithm used as one part of long and complex decision-producing chain
Ensure intended and actual outcomes are fair and transparent	✓	✓	✓
Ensure outcomes can be explained	✓	~	~
Ensure there is a process of oversight and evaluation	✓	✓	✓
Ensure a person is accountable for decisions made using the AI	✓	✓	~
Ensure outcomes are fair, inclusive and respect dignity and rights	✓	✓	~
Ensure use of AI is known and there is a process of appeal	✓	~	~
Mitigate against harms	✓	✓	✓

Where does this leave us?

A one size-fits-all approach that covers all instances of AI is unlikely to work, unless it's so high-level as to offer little practical guidance… But that might be ok

Any code that is created needs to cover private sector partners providing services to government

Most public sector applications of AI are closer to level 1 than level 3, so implementing a code is possible… for now

The biggest unresolved issue is "explainability" - should we choose to avoid uses of AI we cannot adequately explain?

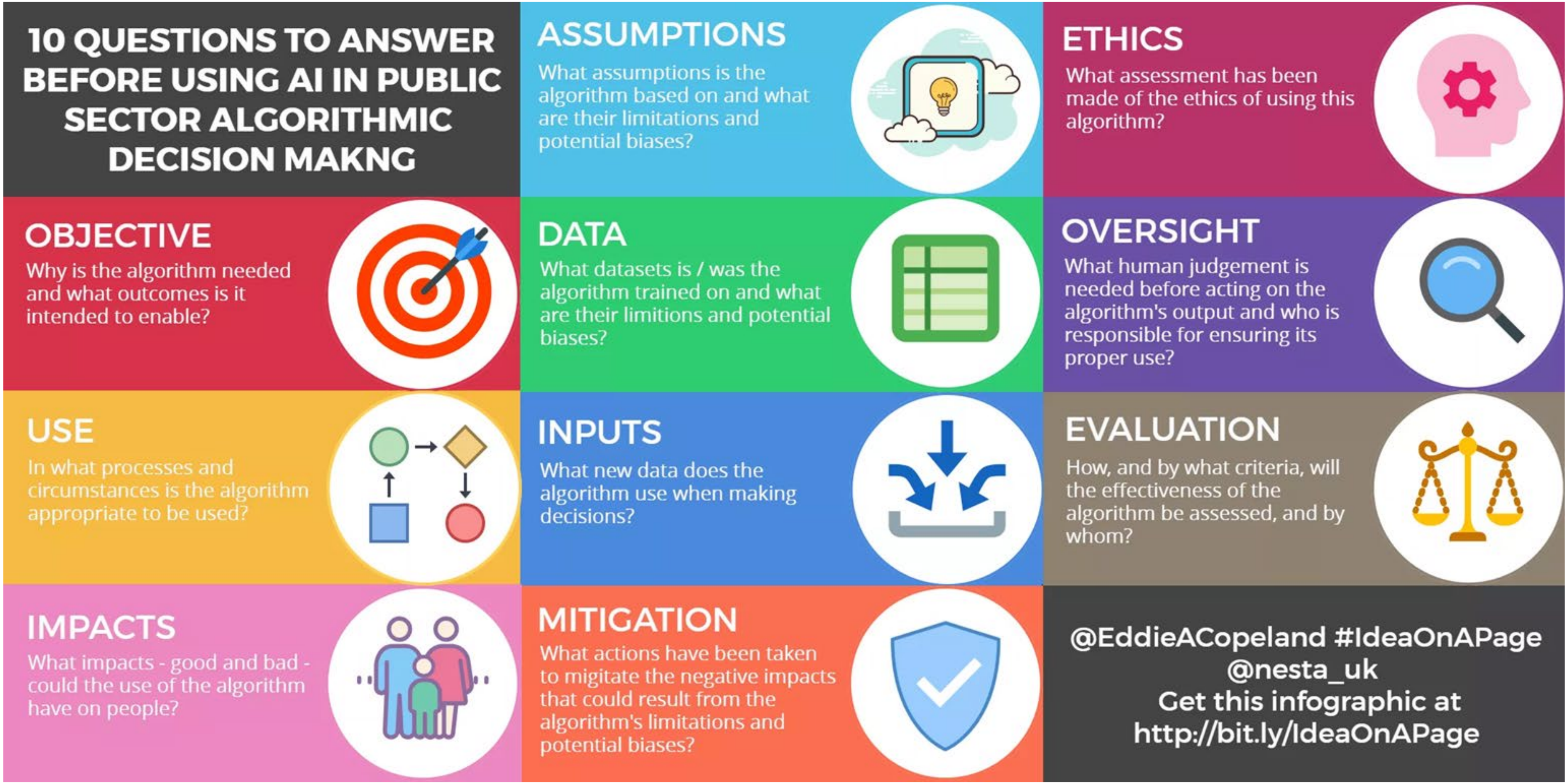
“It will be possible to assess a predictive algorithm’s politics, performance, fairness, and relationship to governance only with significant transparency about how the algorithm works.”

Algorithmic Transparency for the Smart City

We need a diverse set of people involved at every stage of design, oversight and evaluation of AI

Can we have more emphasis on - and faith in - professional judgement?

Nesta’s 10 questions for public sector use of AI



10 QUESTIONS TO ANSWER BEFORE USING AI IN PUBLIC SECTOR ALGORITHMIC DECISION MAKNG

ASSUMPTIONS

What assumptions is the algorithm based on and what are their limitations and potential biases?

ETHICS

What assessment has been made of the ethics of using this algorithm?

OBJECTIVE

Why is the algorithm needed and what outcomes is it intended to enable?

DATA

What datasets is / was the algorithm trained on and what are their limitions and potential biases?

OVERSIGHT

What human judgement is needed before acting on the algorithm's output and who is responsible for ensuring its proper use?

USE

In what processes and circumstances is the algorithm appropriate to be used?

INPUTS

What new data does the algorithm use when making decisions?

EVALUATION

How, and by what criteria, will the effectiveness of the algorithm be assessed, and by whom?

IMPACTS

What impacts - good and bad - could the use of the algorithm have on people?

MITIGATION

What actions have been taken to migitate the negative impacts that could result from the algorithm's limitations and potential biases?

@EddieACopeland #IdeaOnAPage
@nesta_uk
Get this infographic at
http://bit.ly/IdeaOnAPage

Education and awareness about these issues is needed for all those who will work with AI