Compulsory HLT

1. **Find out what Responsible AI is?**

Responsible AI is really all about the how: how do we design, develop and deploy these systems that are fair, reliable, safe and trustworthy. And to do this, we need to think of Responsible AI as a set of socio-technical problems. We need to go beyond just improving the data and models. We also have to think about the people who are ultimately going to be interacting with these systems.”

**Fairness**

The fairness of AI systems is crucially important now that AI plays an increasing role in our daily lives. That’s why Microsoft researchers are advancing the frontiers of research on this topic, focusing on many different aspects of fairness, including:

**Definitions:** Different types of [fairness-related harms](https://www.youtube.com/watch?v=fMym_BKWQzk) that occur in the context of AI, including [harms that are specific to people with disabilities](https://www.microsoft.com/en-us/research/publication/toward-fairness-in-ai-for-people-with-disabilities-a-research-roadmap/) and harms arising from [data quality issues, methodological pitfalls, and ethical limitations](https://www.microsoft.com/en-us/research/publication/social-data-biases-methodological-pitfalls-and-ethical-boundaries/).

**Development practices:** Ways to make fairness a priority throughout the AI development and deployment lifecycle by [identifying industry practitioners’ needs](https://www.microsoft.com/en-us/research/publication/improving-fairness-in-machine-learning-systems-what-do-industry-practitioners-need/) for support in developing fairer AI systems and understanding [organizational challenges and opportunities](https://www.microsoft.com/en-us/research/publication/co-designing-checklists-to-understand-organizational-challenges-and-opportunities-around-fairness-in-ai/) around fairness in AI.

**Applications:** Fairness-related harms in natural language processing and [information retrieval](https://www.microsoft.com/en-us/research/publication/facts-ir-fairness-accountability-confidentiality-transparency-and-safety-in-information-retrieval/), such as [gender stereotypes reflected in word embedding’s](https://www.microsoft.com/en-us/research/publication/quantifying-reducing-stereotypes-word-embeddings/), [problematic predictive text](https://www.meetup.com/SEA-Search-Engines-Amsterdam/events/tpklmrybccbpc/), and [homogenous search results](https://www.microsoft.com/en-us/research/publication/algorithmic-greenlining-an-approach-to-increase-diversity/), as well as ways to [leverage some of these harms to achieve fairer outcomes](https://www.microsoft.com/en-us/research/publication/whats-in-a-name-reducing-bias-in-bios-without-access-to-protected-attributes/) in other tasks.

**The law:** The relationship between AI and the law, including tensions between antidiscrimination laws and the use of AI systems in employment from both a [disparate treatment](https://www.microsoft.com/en-us/research/publication/stretching-human-laws-to-apply-to-machines-the-dangers-of-a-colorblind-computer/) and a [disparate impact](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899) perspective.

**Transparency and Intelligibility**

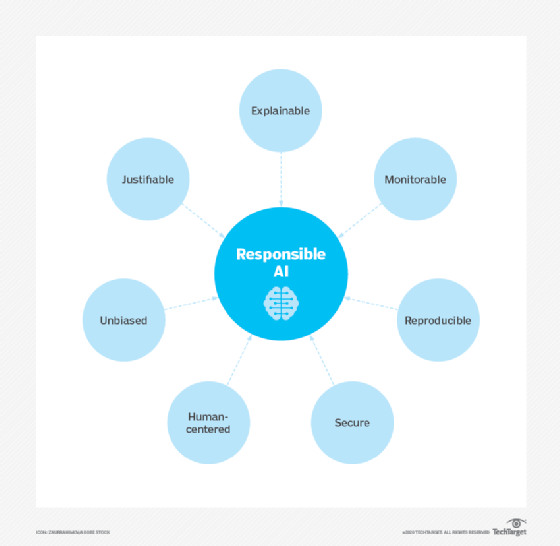
Intelligibility can uncover potential sources of unfairness, help users decide how much trust to place in a system, and generally lead to more usable products. It also can improve the robustness of machine learning systems by making it easier for data scientists and developers to identify and fix bugs. Because intelligibility is a fundamentally human concept, it’s crucial to take a [human-centred approach](http://www.jennwv.com/papers/intel-chapter.pdf) to designing and evaluating methods for achieving intelligibility.  That’s why Microsoft researchers are [questioning common assumptions](https://www.microsoft.com/en-us/research/publication/manipulating-and-measuring-model-interpretability/) about what makes a model “interpretable,” studying [data scientists’ understanding and use](http://www.jennwv.com/papers/interp-ds.pdf) of existing intelligibility tools and how to make these tools [more useable](https://www.microsoft.com/en-us/research/publication/gamut-a-design-probe-to-understand-howdata-scientists-understand-machine-learning-models/), and exploring the intelligibility of [common metrics like accuracy](https://www.microsoft.com/en-us/research/publication/understanding-the-effect-of-accuracy-on-trust-in-machine-learning-models/).

**Reliability and Safety**

Reliability is a principle that applies to every AI system that functions in the world and is required for creating trustworthy systems. A reliable system functions consistently and as intended, not only in the lab conditions in which it is trained, but also in the open world or when they are under attack from adversaries. When systems function in the physical world or when their shortcomings can pose risks to human lives, problems in system reliability translate to risks in safety.

**Human-AI Interaction and Collaboration**

Advances in AI have the potential to enhance human capabilities and improve our lives. At the same time, the complexities and probabilistic nature of AI-based technologies presents unique challenges for safe, fair, and responsible human-AI interaction. That’s why Microsoft researchers are taking a human-centred approach to ensure that what we build benefits people and society, and that how we build it begins and ends with people in mind.



1. **Find instances where AI has failed? Or been used maliciously or incorrectly.**

**AI fails to do image recognition**

Deep learning, the set of algorithms that is often used to implement AI, started its triumphal procession with the breakthrough in image recognition, also known as Computer Vision about 20 years ago. It solved earlier unsolvable task of distinguishing cats from dogs and vice-versa, and went on with more complex and demanding tasks. Now it is a common possession to believe that the computer vision is a robust and reliable technology that can hardly fail.

However, a year ago, researchers from Berkeley, University of Chicago and University of Washington collected 7,500 unedited nature photos which confuse the most advanced computer vision algorithms.

**AI for secure system access by face can be fooled with a mask**

If you have an iPhone X with Face ID, make sure no one has a mask with your face. Apple said that Face ID used the iPhone X’s advanced front-facing camera and machine learning to create a 3-dimensional map of your face. The machine learning/AI component helped the system adapt to cosmetic changes (such as putting on make-up, donning a pair of glasses, or wrapping a scarf around your neck), [without compromising on security](https://support.apple.com/en-ca/HT208108).

Vietnam-based security firm [Bkav found](http://www.bkav.com/dt/top-news/-/view_content/content/103968/bkav%EF%BF%BDs-new-mask-beats-face-id-in-twin-way-severity-level-raised-do-not-use-face-id-in-business-transactions) that they could successfully unlock a Face ID-equipped iPhone by gluing 2D “eyes” to a 3D mask. The mask, made of stone powder, cost around $200. The eyes were simple, printed infrared images.

**Fail: Amazon Axes their AI for Recruitment Because Their Engineers Trained It to be Misogynistic**

Artificial intelligence and machine learning have a huge bias problem. Or rather, they have a huge problem with bias. And the launch, drama, and subsequent ditching of Amazon’s AI for recruitment is the perfect poster-child.

Amazon had big dreams for this project. As one Amazon engineer [told The Guardian in 2018](https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine), “They literally wanted it to be an engine where I’m going to give you 100 résumés, it will spit out the top five, and we’ll hire those.”

But eventually, the Amazon engineers realized that they’d taught their own AI [that male candidates were automatically better](https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine).

How did this AI fail happen? In short, Amazon trained their AI on engineering job applicant résumés. And then they benchmarked that training data set against current engineering employees.

Now, think about who applies for software engineering jobs. And who is most-likely to be currently-employed in software engineering? That’s right: [white men](https://www.seattletimes.com/business/amazon/amazon-more-diverse-at-its-warehouses-than-among-white-collar-ranks/).

So, from its training data, Amazon’s AI for recruitment “learned” that candidates who seemed whiter and more male were more-likely to be good fits for engineering jobs.

1. **Implications of when AI fails. There is a specific article in the GDPR Law that covers this, especially with automated decision making. (opt in and out options).**

The first paragraph of Article 22 provides for a general right not to be subject to completely  
automated decisions significantly affecting the data subject:  
 The data subject shall have the right not to be subject to a decision based solely on  
automated processing, including profiling, which produces legal effects concerning  
him or her or similarly significantly affects him or her.  
Even though this provision refers to a right, it does not provide for a right to object to automated  
decision-making, namely, it does not assume that automated decision-making is in general  
permissible as long as the data subject does not object to it. It rather introduces a prohibition upon  
controllers: automated decisions affecting data subjects are prohibited, unless they fit in one of the  
exceptions provided in paragraph 2.92 According to the Article 29 Working Party:  
 as a rule, there is a general prohibition on fully automated individual decision-making,  
including profiling that has a legal or similarly significant effect.93  
For the application of the prohibition established by Article 22(1), four conditions are needed: a  
decision must be taken, (2) it must be solely based on automated processing, (3) it must include  
profiling, (4) it must have legal or anyway significant effect.  
The first condition requires that a stance be taken toward a person, and that this stance is likely to  
be acted upon (as when assigning a credit score).  
The second condition requires that humans do not exercise any real influence on the outcome of a  
decision-making process, even though the final decision is formally ascribed to a person. This  
condition is not satisfied when the system is only used as a decision-support tool for human beings,

STOA | Panel for the Future of Science and Technology  
60  
who are responsible for the decision, deliberate on the merit of each case, and autonomously decide  
whether to accept or reject the system's suggestions.94  
The third condition requires that the automated processing determining the decision includes  
profiling. A different interpretation could be suggested by the comma that separates 'processing'  
and 'including profiling' in Article 22(1), which seems to indicate that profiling only is an optional  
component of the kind of automated decisions that are in principle prohibited by Article 22(1).  
However, the first interpretation (the necessity of profiling) is confirmed by Recital (71), according  
to which the processing at stake in the regulation of automated decision must include profiling:  
Such processing includes 'profiling' that consists of any form of automated processing  
of personal data evaluating the personal aspects relating to a natural person, in  
particular to analyse or predict aspects concerning the data subject's performance at  
work, economic situation, health, personal preferences or interests, reliability or  
behaviour, location or movements.  
The fourth condition requires that the decision  
produces legal effects concerning[the data subject] or similarly significantly affects him  
or her.  
Recital (71) mentions the following examples of decision having significant effects: the 'automatic  
refusal of an online credit application or e-recruitingpractices'.95 It has been argued that such effects  
cannot be merely emotional, and that usually they are not caused by targeted advertising, unless  
'advertising involves blatantly unfair discrimination in the form of web-lining and the discrimination  
has non-trivial economic consequences (e.g., the data subject must pay a substantially higher price  
for goods or services than other persons).'96  
Many decisions made today by AI systems fall under the scope of Article 21(1), as AI algorithms are  
increasingly deployed in recruitment, lending, access to insurance, health services, social security,  
education, etc. The use of AI makes it more likely that a decision will be based 'solely' on automated  
processing. This is due to the fact that humans may not have access to all the information that is  
used by AI systems, and may not have the ability to analyse and review the way in which this  
information is used. It may be impossible, or it may take an excessive effort to carry out an effective  
review – unless the system has been effectively engineered for transparency, which in some cases  
may be beyond the state of the art. Thus, especially when a large-scale opaque systemic deployed,  
humans are likely to merely execute the automated suggestions by AI, even when they are formally  
in charge. Moreover, human intervention may be prevented by the costs-and-incentives structure  
in place: humans are likely not to substantially review automated decision, when the cost of  
engaging in the review – frogman individual or an institutional perspective– exceeds the significance  
of the decision (according to the decision-maker's perspective).

3.6.2. Article 22(2) GDPR: Exceptions to the prohibition of 22(1)  
Paragraph 2 of Article 22 provides for three broad exceptions to Paragraph 1. It states that the  
prohibition on automated decision-making does not apply when the processing upon which the  
decision is based  
 a) is necessary for entering into, or performance of, a contract between the data  
subject and a data controller;

b) is authorised by Union or Member State law to which the controller is subject, and  
which also lays down suitable measures to safeguard the data subject's rights and  
freedoms and legitimate interests; or  
 c) is based on the data subject's explicit consent.  
Based on the broad exception of item (a), automated decision-making is enabled in key areas such  
as recruitment and lending. However, for the exception to apply, decisions based solely on  
automated processing must be ‘necessary.' Such necessity may depend on the high number of cases  
to be examined (e.g., a very high number of applications to a job). The necessity of using AI in  
decision-making may also be connected to AI capacities to outperform human judgement. In this  
connection we may wonder whether human involvement will still contribute to a stronger  
protection of data subjects, or whether the better performance of machines – even with regard to  
the political and legal values at stake, e.g. Ensuring 'fair equality of opportunity' for all applicants to  
a position97 – will make human intervention redundant or dysfunctional. Outside of the domain of  
contract and legal authorisation, consent may provide a basis for automated decision-making  
according to Article 22(2)(c). However, the conditions for valid consent not always obtain, even in  
cases when automated decision-making seems appropriate. Consider for instance the case in which  
an NGO uses an automated method for classifying(profiling) applicants to determine their need and  
consequently allocate certain benefits to them. In such a case, it is very doubtful that an applicant's  
consent may be viewed as free (as not consenting would entail being excluded from the benefit),  
but the system seems socially acceptable and beneficial even so.

1. **What should organisations do to ensure that they are being responsible with AI and the wider use of data in general?**

When it comes to AI, we have seen an increase in appetite from business leaders to be at the forefront of pioneering AI technologies – from only 14% in 2018, to 28% in 2019. This underlies the increased urgency in getting ahead with AI to enable successful business outcomes, but also created an increased urgency to ensure we get things right with AI. With all the truly amazing progress that has been made in AI over the past year, it is still important to remember that we are at a very early stage of truly understanding the magnitude of impact on our global society should this technology remain unchecked.

To ensure ongoing public trust in their brand, organisations must consider the long-term reputational and cultural benefits of moving beyond just discussing high level principles on the ethical use of AI and focus on what this means in practice when they implement and deploy AI. Regulators have an important role to play here as well – if they take a risk-based approach, that is focused on outcomes rather than technology in order to support innovation, for example anti-discriminatory regulation which is technology agnostic.

## **3 steps to ensure AI is served in a responsible way**

So, where do we go from here to ensure that AI is serving our society in a healthy and responsible way? Organisations must think of AI technology in a holistic way – understanding where AI sits in the value chain and creating the right structures to ensure long-term governance by:

1. Establishing internal governance, for example by an objective review panel, that is diverse and that has the knowledge to understand the possible consequences of AI infused systems. A key success factor is leadership support and the power to hold leadership accountable.
2. Ensuring the right technical guardrails, creating quality assurance and governance to create traceability and auditability for AI systems. This is an important part of every organisation’s toolkit to allow operational and responsible AI to scale.
3. Investing more in their own AI education and training so that all stakeholders – both internal and external – are informed of AI capabilities as well as the pitfalls.