

AI GRCA PRACTITIONER PROGRAMME • MODULE 2

Responsible AI, Trustworthy AI and Harm

Facilitated by Collins Obasuyi

PADEMIQ / AI GRCA Lab · 90-minute session

TODAY'S 90 MINUTES

Session roadmap

0:00	Welcome, aim & practitioner mindset	8 min
0:08	01 · Responsible & Trustworthy AI	10 min
0:18	02 · Fairness & bias	20 min
0:38	*Break*	5 min
0:43	03 · Transparency, explainability & accountability	10 min
0:53	04 · Privacy, safety & security	7 min
1:00	05 · Human oversight & automation bias	7 min
1:07	06 · Understanding AI harm	6 min
1:13	07 · From principle to control	4 min
1:17	**08 · Workshop: HireAI Harm Map**	10 min
1:27	Wrap-up & homework	3 min

SESSION AIM

8 min

What could this AI do to people?

By the end of today you should understand what Responsible AI and Trustworthy AI mean in practice, how AI can create unfair or harmful outcomes, where bias enters the lifecycle — and how to identify the people who may be affected.

MODULE 1

Understand the AI system

Evidence: AI System Description

MODULE 2

Understand its impact on people

Evidence: AI Harm Map

Moving from principles to evidence.

WHAT THIS MODULE IS NOT

Not memorising ethical principles.

We’re not turning anyone into philosophers.

We’re not putting “fairness, transparency, accountability, explainability” on a PowerPoint and declaring the AI responsible.

We’re turning those words into practical questions.

Every claim gets a question

WHEN SOMEONE SAYS…	WE ASK…
“Our AI is fair.”	**How do you know?**
“There is human oversight.”	**What does the human actually do?**
“The system is transparent.”	**Transparent to whom?**
“We have tested the model.”	**What did you test, against what standard — and where is the evidence?**



THE MINDSET

This is the practitioner mindset: every claim about AI should be backed by evidence.

LAST MODULE → THIS MODULE

You cannot govern what you cannot describe.

What happens to people when this system is wrong?

Could it work exactly as designed — and still cause harm?

Yes.

An AI can be fast, cheap, accurate most of the time — even beat humans on a test — and still harm people. That is where Responsible AI begins.

10 min

PART 01

01

Responsible and Trustworthy AI

From “we care about fairness” to “here is the evidence”.

IN THIS PART

01 Responsible AI

02 Principles vs controls

03 Trustworthy AI

04 Principles in context

Using AI without switching off responsibility

DEFINITION

Responsible AI means designing, developing, deploying and using AI in a way that considers its impact on people — and manages the risks properly.

STILL RESPONSIBLE, EVEN WHEN AI RECOMMENDS

- **The organisation**
- **The business owner**
- **The risk owner**
- **The person approving the system**



Responsible AI needs evidence — not just good intentions.

KEY TEACHING POINT

Responsible AI is not a statement. It is something an organisation must be able to demonstrate.

EVERYDAY EXAMPLE

A responsible driver is judged by what they do



The principle

Road safety

THE CONTROLS — WHAT YOU ACTUALLY DO

- ✓ wear a seat belt
- ✓ maintain the brakes
- ✓ check the tyres
- ✓ follow speed limits
- ✓ don’t drive drunk
- ✓ keep insurance
- ✓ complete an MOT
- ✓ pay attention
- ✓ respond to warning lights

AI IS THE SAME

Responsible AI isn’t “we care about fairness”. It’s “we identified fairness risks, tested, documented, built controls, assigned owners and keep monitoring”.

When someone says they use Responsible AI, ask...

1. What principles do you follow?
2. How are they turned into controls?
3. Who owns each control?
4. How do you test those controls?
5. What evidence exists?
6. What happens if a principle is breached?
7. Who decides whether the system can still be used?
8. How often is the system reviewed?



REMEMBER

The principle is the promise. The controls, owners and evidence are the proof.

Trust should come from evidence — not confidence

DEFINITION

Trustworthy AI is AI that people have good reason to trust — because important risks are understood, controlled and monitored.



THE WRONG TYPE OF TRUST

Right 20 times → "it's brilliant" → people stop checking → wrong answer, nobody notices.

YOU TRUST A MECHANIC BECAUSE

- they are qualified and experienced
- they inspect the car
- they explain the problem
- they keep records
- their work can be checked
- someone is accountable if it goes wrong

NOT BECAUSE

"They sounded confident."

AI should be trusted the same way: because it can be checked, not because it sounds clever.

The practice — and the confidence it earns

RESPONSIBLE AI = THE PRACTICE

How we design, govern and use AI responsibly.

TRUSTWORTHY AI = THE CONFIDENCE

Whether there's enough evidence to justify trust.

TRUSTWORTHY AI MAY INVOLVE

- reliability
- fairness
- transparency
- explainability
- accountability
- privacy
- safety
- security
- human oversight
- resilience
- monitoring

KEY TEACHING POINT

Trust is earned through evidence. Not branding. Not vendor promises. Not because the demo looked impressive.

The words change. The governance challenge doesn't.

fairness | transparency | explainability | accountability | privacy | safety | security | human oversight | reliability | robustness

For every principle, ask: what does this mean for this specific system?

Fairness for	Netflix	*is not the same as for*	**Recruitment**
Safety for	A writing assistant	*is not the same as for*	**Medical AI**
Explainability for	A music recommendation	*is not the same as for*	**A loan rejection**



KEY TEACHING POINT

Principles are broad. Governance turns them into specific requirements — because context matters.

20 min

PART 02

02

Fairness and bias

How AI creates unfair outcomes — often without anyone intending it — and where bias gets in.

IN THIS PART

01 What fairness means

02 What bias is

03 Six ways bias gets in

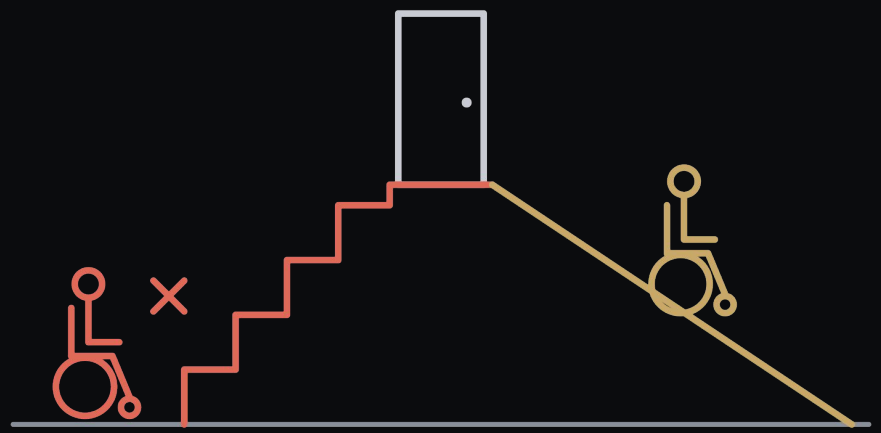
04 Feedback loops & the lifecycle

05 When accuracy hides unfairness

Are some people getting a worse outcome?

DEFINITION

Fairness means making sure an AI system does not create unjustified disadvantage for people or groups.



EVERYDAY EXAMPLE — STAIRS & WHEELCHAIR ACCESS

"We treat everyone equally. Everyone uses the stairs."

It sounds equal — but someone using a wheelchair cannot get in. Treating everyone identically created an unfair outcome. The ramp recognises different needs.

KEY TEACHING POINT

Fairness sometimes means recognising different needs or circumstances — not giving everyone the same score.

A neutral-looking feature can have an unfair effect

HireAI scores candidates partly on continuous employment history. But career breaks happen because of:

- disability
- illness
- caring responsibility
- pregnancy
- military service
- unemployment
- education
- family circumstances

The feature looked neutral.

The effect may not be neutral.

Fairness questions: Are outcomes or error rates different across groups? Are some people rejected more often? Could a variable act as a proxy? Can people challenge the outcome?



KEY TEACHING POINT

A system can create unfair outcomes even when nobody intended to discriminate.

The system keeps leaning in one direction



Bias is a systematic pattern that can cause an AI system to favour, disadvantage or misrepresent certain people, groups or situations.

EVERYDAY

Restaurant reviews

90 of 100 reviews come from one type of customer: “Everyone loves spicy food.” Maybe not.

AI EXAMPLE

Voice recognition

Trained mostly on one accent: great for that group, worse for others.

BIAS CAN ENTER THROUGH

- data
- labels
- design
- assumptions
- measurement
- deployment
- human behaviour
- feedback loops

KEY TEACHING POINT

Bias can exist even when overall system performance looks good.

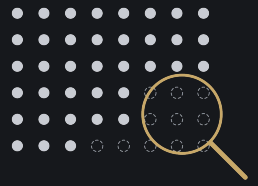
A quick map before we zoom in



1

Historical bias

Past data reflects unfair patterns — the AI learns yesterday's mistakes.



2

Representation bias

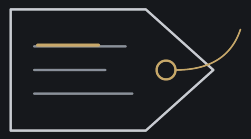
Some people or situations are missing or underrepresented.



3

Measurement bias

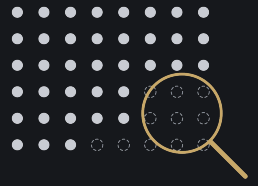
We measure an easy proxy, not the thing we actually care about.



4

Labelling bias

Labels are somebody's opinion stored as data.



5

Selection bias

The data isn't a fair or complete sample — the missing data matters most.



6

Proxy discrimination

Neutral-looking data (postcode, graduation year) reveals sensitive traits.

Historical data can teach AI yesterday's mistakes



Past data reflects unfair or unequal patterns that already existed — and the AI learns from that history.

10 years of data

A company hired mostly men for senior technical roles.

→

The model sees

More men hired, promoted and rated highly.

→

The AI may learn

"Profiles like those men are successful."

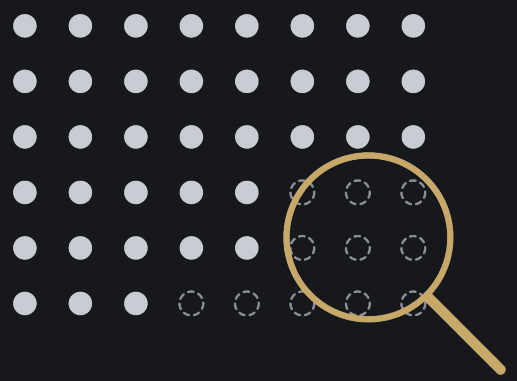
LENDING TOO

If some communities historically had less access to financial services, a model trained on old lending outcomes may learn that pattern — even without ever using race directly.

KEY TEACHING POINT

Historical data is not automatically objective. History contains human decisions — and human decisions contain human bias.

The AI hasn’t seen enough examples of everyone



Representation bias: some people, groups, environments or situations are underrepresented in the data.

EVERYDAY

Learning dogs

- Train someone only on German Shepherds, Labradors and Golden Retrievers.
- Then show them a Chihuahua.
- They struggle — the examples were too narrow.
- Same with facial recognition trained mostly on one demographic group.

HIREAI

Trained in one place, used everywhere

- Trained on applications from one large London tech company.
- Later sold to hospitals, councils, retailers, construction firms and charities.
- Different workforces, roles and applicants — the data may not represent the new use.

KEY TEACHING POINT

A model may work well for the people it has seen — and poorly for the people it has not.

Measuring the easy thing instead of the real thing



Measurement bias: what we measure isn't a good representation of what we actually care about.

EVERYDAY

"Healthy = 10,000 steps a day"

- Easy to measure — but health is much bigger than steps.
- People swim, cycle, lift weights, use wheelchairs, do manual work or live with a condition.
- The measure is incomplete.

HIREAI

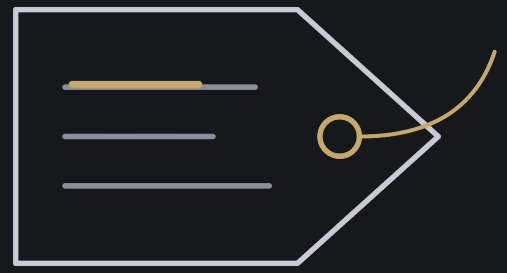
"Good employee" = performance rating

- "Good employee" is hard to measure, so they use annual ratings.
- Ratings can reflect manager bias, team culture, favouritism, visibility and unequal opportunity.
- The AI learns from the measurement — flaws included.

KEY TEACHING POINT

AI can calculate perfectly using a bad measurement. The maths can be correct while the decision is still wrong.

Labels are not automatically truth



Labelling bias: labels used for training are inaccurate, inconsistent or shaped by human judgement.

EVERYDAY

Spam or not spam?

- People label emails by hand.
- One labels marketing emails as spam; another as legitimate.
- Inconsistent labels — and the AI learns from them.

HIREAI

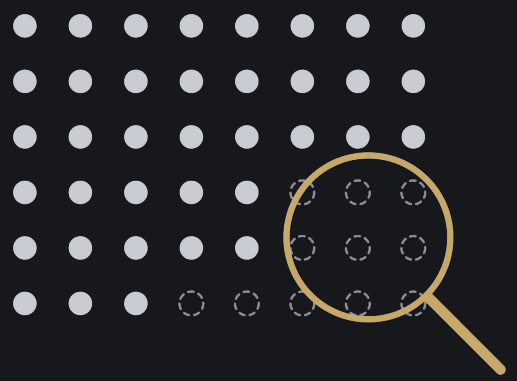
“Good culture fit”

- Records say strong / weak candidate, good / poor culture fit.
- Who decided what “culture fit” meant? Was it consistent?
- Biased human judgements can become training labels.

KEY TEACHING POINT

Sometimes labels are somebody’s opinion stored as data.

Sometimes the missing data matters most



Selection bias: the data used is not a fair or complete sample of the population.

EVERYDAY

“95% of customers are satisfied”

- Only customers who completed an online survey were asked.
- Unhappy customers may never have filled it in.
- The sample doesn’t tell the whole story.

HIREAI

Learning only from people who were hired

- What about the candidates who were rejected?
- Some may have been excellent — but their future performance was never observed.
- The training data has a built-in blind spot.

KEY TEACHING POINT

Sometimes the missing data is more important than the data you have.

Removing “age” doesn’t remove age



A variable that looks neutral indirectly reveals — or closely relates to — a sensitive characteristic.

Doesn’t use: ethnicity

BUT IT USES

- postcode

Postcode may correlate with ethnicity, income, deprivation, housing and education.

Doesn’t use: age

BUT IT USES

- graduation year
- length of career
- years of experience

Together these may indirectly reveal age. No “Age” field needed.

Doesn’t use: gender

BUT IT USES

- memberships
- career breaks
- job titles
- education
- hobbies

Some combinations may reveal gender or other characteristics.

KEY TEACHING POINT

Removing a protected characteristic does not automatically remove discrimination risk.

The AI creates the evidence that seems to prove it right



An AI's output affects the future data it later learns from.

POLICING EXAMPLE

Predicts more crime in an area → More police sent → More incidents detected → "See? More crime here."

HIREAI EXAMPLE

Recommends fewer from one background → Fewer hired → Even fewer examples in the data → Next model learns it more strongly

Ask: does the model influence the data collected later? Are rejected cases ever reviewed? Are outcomes monitored independently of the model?

KEY TEACHING POINT

AI systems can reinforce their own mistakes if feedback loops are not monitored.

Bias can enter at every stage — not just in the data

1 Idea

“Let’s predict which employees might leave.” Should we predict this at all?

2 Use case

Built for workforce planning, later used for promotions. Use changed; risk changed.

3 Data

Missing groups, historical bias, poor quality, proxies, bad labels.

4 Design

Assumes continuous employment means reliability.

5 Development

Feature, model and optimisation choices shape outcomes.

6 Testing

Only average performance tested — nobody checks across groups.

7 Deployment

Used on a different population from the one it was trained on.

8 Human use

One recruiter challenges; another follows every recommendation.

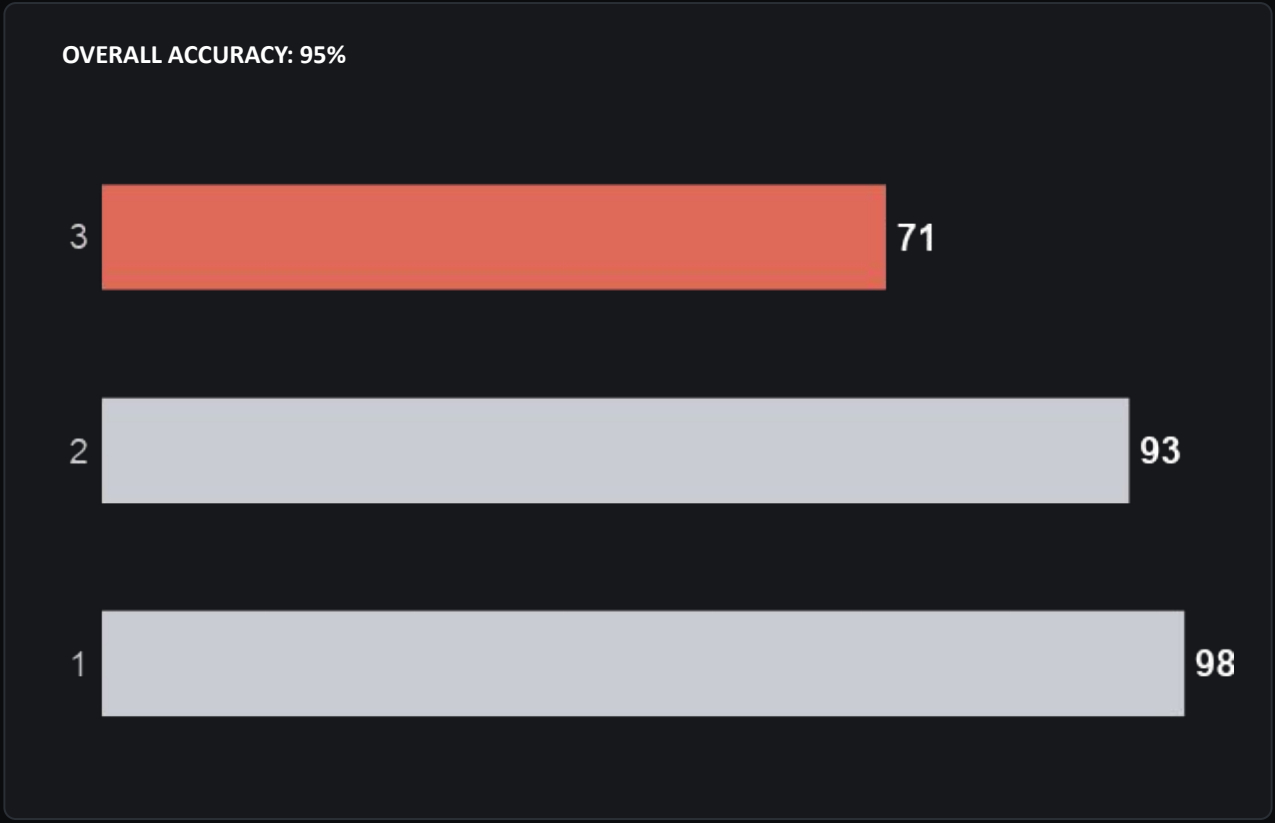
9 Monitoring

Without it, unfair patterns can continue for years.

KEY TEACHING POINT

Bias is a lifecycle problem. Not just a data problem.

95% accurate — for whom?



EVERYDAY

“95% of our students passed.”

Then: one class 99%, another 94% — another 50%. The average hides the problem.

ASK

- How does performance vary between groups?
- What are the false positive and false negative rates?
- Are errors distributed fairly?
- Which errors cause the most harm?

KEY TEACHING POINT

Don’t stop at average performance. Ask who experiences the errors.

Different mistakes, different harms

FALSE POSITIVE

Says something is true when it isn't

Fraud system: "Fraud." — but the transaction was genuine.

FALSE NEGATIVE

Says something isn't true when it is

Fraud system: "Not fraud." — but it was fraud.

RECRUITMENT

"Not suitable" — but they'd have been excellent

Another candidate may be rated highly but perform poorly later. Different errors, different harms.

MEDICAL

"No urgent problem" — but it's serious

Here a false negative could cause severe harm.

KEY TEACHING POINT

Accuracy doesn't tell us everything. Understand what kind of mistakes the AI makes — and who carries the harm.

The practitioner’s bias checklist

1. What historical period does the data represent — were past decisions fair?
2. Who is represented in the data — and who is missing?
3. Does the training population match the deployment population?
4. What are we predicting, and is it measured fairly?
5. Who created the labels — were they checked?
6. Could any feature act as a proxy for a protected characteristic?
7. Does the model influence the data it later learns from?
8. Has performance been tested across groups?
9. Are error rates different across groups?
10. Can affected people challenge the outcome?

5 min



5-minute break

When we return: transparency, explainability, accountability — and what AI could do to people.

10 min

PART 03

03

Transparency, explainability and accountability

Can people see the AI, understand its outputs — and find who is responsible?

IN THIS PART

01 Transparency

02 Explainability

03 The difference

04 Accountability

People shouldn’t be kept in the dark about important AI use

DEFINITION

Transparency means providing appropriate information about an AI system, its purpose, its use and its limitations.

TRANSPARENCY MAY INCLUDE

- telling people AI is being used
- explaining the purpose and the AI’s role
- explaining what data is used
- explaining limitations
- identifying who is responsible
- explaining how to challenge an outcome



HIREAI

Rejected — never knew AI was involved

Is that transparent? Probably not enough.



CHATBOT

They think they’re talking to a human

The organisation may need to make it clear it’s AI.



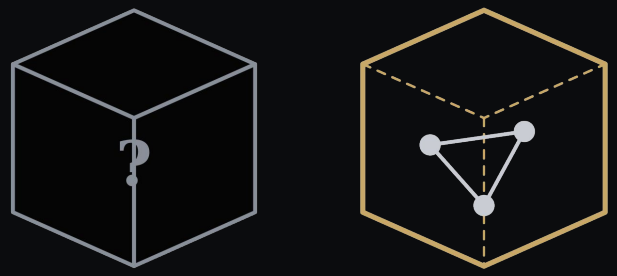
KEY TEACHING POINT

Transparency isn’t information dumping — or revealing source code and secrets. It’s giving the right people (customer, regulator, auditor, affected person) what they need.

Why did the AI say that?

DEFINITION

Explainability means being able to give a meaningful explanation of how or why an AI output or decision was reached.



Black box vs glass box

NOT AN EXPLANATION

Score: 42

"Why did I score 42?" — "The algorithm scored you 42" is not an answer. Nor is "the gradient boosted ensemble returned 0.62".

A MEANINGFUL EXPLANATION

- A required qualification was missing
- Required experience wasn't identified
- Responses didn't meet defined criteria

KEY TEACHING POINT

An explanation is only useful if the person receiving it can understand it — and more important as stakes rise (song → loan → diagnosis).

About the system — or about this output?

TRANSPARENCY

Tells you about the system

- Is AI being used?
- What is it used for?
- What data does it use?
- Who owns it? What are its limits?

EXPLAINABILITY

Helps you understand an output

- Why did this output happen?
- What factors influenced it?
- Can the decision be challenged?

"We use AI to support recruitment screening." → **transparency**

"Why was my application scored lower?" → **explainability**

KEY TEACHING POINT

Transparency tells you about the system. Explainability helps you understand a specific output or decision.

Never accept “the AI did it”

DEFINITION

Accountability means someone is responsible for the AI system, its decisions, its controls and its consequences.

VENDOR

“The employer controls recruitment.”

→

EMPLOYER

“It’s the vendor’s technology.”

→

RECRUITER

“I only follow the recommendation.”

Ask: Who owns the system, the process and the risk? Who approves deployment, owns fairness testing, monitors, handles complaints and incidents — and who can pause it?

KEY TEACHING POINT

If everyone is responsible, nobody is responsible. Accountability must be named.

7 min

PART 04

04

Privacy, safety and security

Just because AI can use more data, act faster or reach more systems doesn’t mean it should.

IN THIS PART

01 Privacy

02 Safety

03 Security

Useful tools can capture more than you expect

AI MEETING ASSISTANT MAY CAPTURE

- employee names
- customer information
- medical information
- commercial secrets
- passwords spoken aloud
- HR discussions

HIREAI MAY PROCESS

- names
- addresses
- employment history
- qualifications
- nationality indicators
- disability information
- career gaps

ASK

1. What personal data is used — and why?
2. Does the organisation need all of it?
3. Who can access it, and for how long?
4. Is it reused, shared or used to train another model?

KEY TEACHING POINT

Just because AI can use more data does not mean it should.

What happens when this system goes wrong?

DEFINITION

AI safety means preventing unacceptable harm caused by the operation, failure or misuse of an AI system.

LOW IMPACT

Netflix recommends the wrong film.

You ignore it.

HIGHER IMPACT

HireAI wrongly rejects a candidate.

They lose an opportunity.

VERY HIGH IMPACT

A medical AI misses a serious condition.

Health or life may be affected.

Ask: What is the worst credible failure? Who could be harmed, and how badly? Can it be detected, reversed or stopped? Is there human intervention and a safe fallback?

KEY TEACHING POINT

A 5% error rate in music recommendations is very different from 5% in medical decision support. Safety is about what happens when it fails.

More access and autonomy means more at stake



Protecting AI systems, models, data and connected tools from unauthorised access, manipulation or attack.

AI SECURITY RISKS MAY INCLUDE

- unauthorised access
- data leakage
- prompt injection
- model manipulation
- data poisoning
- malicious inputs
- stolen credentials
- excessive permissions
- compromised suppliers
- insecure APIs

Chatbot: wrong paragraph → **Assistant: may leak data** → **Agent: sends, updates, deletes** → **Compromised agent: acts for an attacker**

Would you give a new employee your email, bank, CRM and customer database on day one? "They seem smart" isn't access control.

KEY TEACHING POINT

The more access and autonomy AI has, the greater the potential security impact.

7 min

PART 05

05

Human oversight and automation bias

A human in the loop is only a control if that human can actually challenge the AI.

IN THIS PART

01 Meaningful oversight

02 500 clicks & 12 seconds

03 Automation bias

The important word is “meaningful”

DEFINITION

Human oversight means a person has a meaningful ability to review, question, intervene in or override an AI system.

WHAT IT IS NOT

- ✕ *“Someone clicked approve.”*
- ✕ *“A manager can technically override it.”*
- ✕ *“There’s a human somewhere in the process.”*

MEANINGFUL OVERSIGHT MEANS THE PERSON

- understands what the AI does and its limits
- has enough information and time
- has authority to override
- knows when to escalate and what evidence to check
- knows what happens when the AI is wrong

KEY TEACHING POINT

A human in the loop is not automatically a control. The human must actually be able to challenge the AI.

500 clicks. 12 seconds each.



“Every decision is reviewed by a recruiter.”

WEAK OVERSIGHT

Score 38 — Reject

- 500 applications in the queue
- Recruiter clicks “Accept recommendation” 500 times
- Average review: 12 seconds
- Technically a human in the loop — practically, the AI decided

STRONGER OVERSIGHT

Equipped to challenge

- score plus key factors
- comparison with job criteria
- uncertainty warnings
- can inspect the CV and override
- enough time; trained on limits
- overrides recorded and monitored

GOVERNANCE QUESTION

Ask: what exactly does the human review, can they disagree, are overrides recorded — and do they have enough time?

“The computer said it, so it must be right”

DEFINITION

Automation bias happens when people trust an automated recommendation too much, simply because it came from a system.

SAT NAV

Turns into a road they can see is closed

RECRUITER

“The AI probably knows something I don’t.”

CLINICIAN

Ignores their concern because AI says “low risk”

HOW TO REDUCE IT

- user training
- meaningful explanations
- uncertainty indicators
- override processes
- audit of human decisions
- monitoring override rates
- no misleading confidence scores

KEY TEACHING POINT

Human oversight can fail if the human trusts the AI too much.

6 min

PART 06

06

Understanding AI harm

Who gets hurt, disadvantaged, misled, excluded or treated unfairly if this AI goes wrong?

IN THIS PART

01 Types of harm

02 Model failure → human impact

03 Direct & indirect harm

04 Stakeholders & higher-impact groups

Harm does not require bad intention

DEFINITION

AI harm is a negative consequence caused or contributed to by the design, development, deployment or use of an AI system.

- **Financial**
- **Opportunity**
- **Physical**
- **Psychological**
- **Privacy**
- **Reputational**
- **Discriminatory**
- **Safety**
- **Loss of autonomy**
- **Exclusion**
- **Societal**

Who gets hurt?

KEY TEACHING POINT

Harm can happen at individual, group, organisational and societal level.

Always translate model failure into human impact

FINANCIAL HARM

Card blocked while travelling

Fraud AI blocks a genuine card abroad. They can't pay for the hotel, buy food or travel home. Also: loans wrongly rejected, benefits wrongly stopped, pay wrongly denied.

OPPORTUNITY HARM

Qualified candidate never interviewed

Loss of access to a job, education, credit, housing, insurance, healthcare or public services.

THE HARM IS NOT JUST

"The model was inaccurate."

THE HARM IS

"The person lost an employment opportunity."

KEY TEACHING POINT

The model made a prediction. The person experienced the harm.

Bodies, minds, data, reputations — and society

Physical

Unsafe medical advice; a vehicle that misses a pedestrian; equipment controlled incorrectly.

Psychological

Distress, manipulation, emotional dependency — e.g. a wellbeing chatbot responding badly to a vulnerable user.

Privacy

Sensitive data exposed, reused or inferred; profiling people didn't expect; surveillance.

Reputational

Wrongly labelled fraudulent, high risk or unsuitable — damage can remain after correction.

Discriminatory

Unjustified disadvantage linked to age, disability, sex, race, religion or pregnancy.

Societal

Misinformation, discrimination at scale, surveillance, eroded trust, reinforced stereotypes.

KEY TEACHING POINT

Higher physical, psychological or societal risk normally requires stronger controls.

Don’t stop at the system user. Follow the impact.



Direct: HireAI rejects your application.
Indirect: an AI workforce tool cuts someone’s hours — their family feels it too.

HIREAI STAKEHOLDERS

Users	recruiters, hiring managers
Affected people	job applicants
Business	HR director, recruitment lead, managers
Technical	developers, vendor, cloud provider, data team
Governance	risk, compliance, legal, privacy, AI governance, internal audit

KEY TEACHING POINT

The person using the AI is not always the person carrying the risk.

Risk is not experienced equally

MAY EXPERIENCE MORE SERIOUS CONSEQUENCES

- children
- elderly people
- disabled people
- people in financial hardship
- limited digital access
- language barriers
- dependent on essential public services

EXAMPLE

HireAI rejects a candidate — and the only appeal route is “use our AI chatbot”. That may create further barriers for some people. Governance should consider accessibility.

SEVERITY DEPENDS ON CONTEXT

Wrong song recommended	**Low**
Wrong job candidate	**Significant**
Unsafe medical advice	**Severe**
Agent sends wrong internal email	**Embarrassing**
Agent pays the wrong account	**Serious**

KEY TEACHING POINT

Don’t assess AI risk by the technology alone — look at what happens after the output, and to whom.

4 min

PART 07

07

From principle to control

Naming a principle isn’t enough. Practitioners turn it into a control — and prove it works.

IN THIS PART

01 Principle → risk → control → evidence

02 The governance chain

03 Prevent, detect, correct

Turning principles into controls — and evidence

Principle	Risk	Control	Evidence
Fairness	HireAI may unfairly reject candidates from certain groups	Fairness testing completed before deployment	Fairness test report
Transparency	Candidates don’t know AI is used	Recruitment privacy & AI notice explains automated screening	Candidate notice and published recruitment information
Human oversight	Recruiters automatically accept AI recommendations	Rejections require documented recruiter review	Decision logs and override records



Principle → Risk → Control → Test → Evidence → Assurance



THE POINT

A principle without a control is a slogan. A control without evidence is an assumption.

Prevent it. Detect it. Correct it.

PREVENTIVE

Stop it going wrong

- data quality checks
- fairness testing
- prohibited data rules
- approval process & human review
- supplier due diligence
- access control
- pre-deployment testing
- model documentation

DETECTIVE

Spot problems

- monitoring & dashboards
- complaints
- audit logs
- fairness monitoring
- incident reports
- drift monitoring
- anomaly detection

CORRECTIVE

Fix problems

- retraining or updating data
- changing thresholds
- removing a feature
- correcting decisions
- improving human review
- suspending the system
- notifying affected people

KEY TEACHING POINT

Good AI governance normally needs more than one type of control.

10 min

PART 08

08

Workshop: the HireAI Harm Map

Now we build today’s evidence item — together.

IN THIS PART

01 Brief & questions

02 Example harm map

03 The 14-field AI Harm Map

04 Unknowns are findings

6 min

Map the harms HireAI could cause

USING HIREAI, ANSWER

1. Who uses the system?
2. Who is directly affected?
3. Who may be indirectly affected?
4. What could go wrong?
5. Why could it happen?
6. Who experiences the harm?
7. What type of harm is it?
8. How serious could it be?
9. What control could reduce it?
10. What evidence would show it works?
11. What do we still not know?

DON'T WRITE

"Potential disparate impact arising from algorithmic optimisation bias affecting protected subpopulations."

WRITE

"Candidates from some groups may be rejected more often because the model learned from biased hiring history."

RULE

Clear language makes better governance.

4 min

Example HireAI harm map (1 of 2)

	1 · Unfair rejection	2 · Career gaps penalised	3 · No meaningful review
Affected	Job applicant	Candidate with employment gaps	All applicants
What could go wrong?	Qualified candidate rejected	Scored lower because of gaps	Recruiters auto-accept AI rejections
Possible cause	Historical hiring data contains bias	Assumes continuous employment = stronger performance	High workload and automation bias
Type of harm	Opportunity / fairness	Opportunity / discrimination risk	Opportunity / accountability
Possible control	Fairness testing before deployment	Feature review and fairness testing	Mandatory documented review before rejection
Evidence	Fairness evaluation report	Feature documentation and test results	Review logs and override statistics
Unknown	Has the vendor done group-level fairness testing?	How heavily continuity affects the score	Current average recruiter review time

Example HireAI harm map (2 of 2)

	4 · Can’t challenge outcome	5 · Privacy exposure
Affected	Rejected applicant	Job applicant
What could go wrong?	Rejection with no route to question it	Vendor uses data for another purpose
Possible cause	No appeal process	Weak contractual or technical controls
Type of harm	Procedural fairness / opportunity	Privacy
Possible control	Human appeal or reconsideration process	Data-use restrictions and supplier review
Evidence	Appeal procedure and case logs	Contract, data-processing docs, configuration
Unknown	Whether appeals are tracked	Whether candidate data trains the model

KEY TEACHING POINT

The unknowns aren’t failures in the exercise. They are findings.

Each unknown points to a transparency, evidence, ownership, supplier or testing gap.

The AI Harm Map: 14 fields

MODULE 1 ANSWERED

What is this AI system?

MODULE 2 ANSWERS

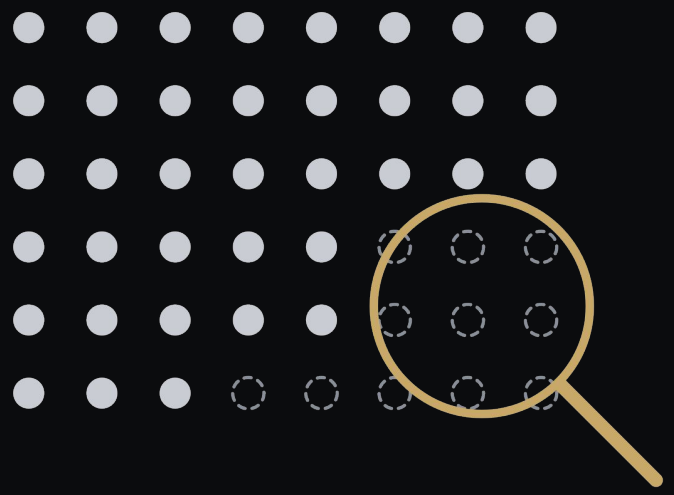
Who could this AI harm, how could that harm happen, and what controls should exist?

1. AI system
2. Stakeholder / affected person
3. Potential harm
4. Harm type
5. Possible cause
6. Bias source
7. Fairness concern
8. Severity
9. Existing control
10. Additional control needed
11. Human oversight
12. Control owner
13. Evidence required
14. Unknowns / gaps

HireAI Harm Map — unfair rejection

1	**AI system**	HireAI — CV Screening Assistant	
2	**Affected person**	Job applicants	
3	**Potential harm**	A qualified candidate may be rejected before interview	
4	**Harm type**	Opportunity harm; potential fairness and discrimination harm	
5	**Possible cause**	Historical hiring decisions may contain bias; the model may favour candidates similar to past hires	
6	**Bias source**	Historical, representation and labelling bias; proxy variables; recruiter behaviour	
7	**Fairness concern**	Some groups may get systematically lower scores or higher rejection rates	
8	**Severity**	High — a candidate may lose access to employment	
9	**Existing control**	Human recruiter reviews AI recommendations	
10	**Additional control**	Group-level fairness testing, feature review, appeal process, recruiter training, outcome monitoring, documented overrides	
11	**Human oversight**	Inspect candidate information, understand key reasons, challenge low scores, override and document	
12	**Control owner**	Recruitment lead, HR risk owner, AI governance lead or model validation — must be confirmed	
13	**Evidence required**	Fairness report, feature docs, training records, decision & override logs, notices, appeal records, dashboard	
14	**Unknowns / gaps**	Training data make-up, group performance, vendor updates, gap weighting, override and appeal rates	

Unknowns are not failures — they are findings



AN UNKNOWN CAN REVEAL

- transparency gaps
- evidence gaps
- ownership gaps
- supplier gaps
- testing gaps
- governance gaps

If you don’t know — write UNKNOWN. Don’t guess.

KEY TEACHING POINT

An unknown can be a governance finding.

3 min

PART 09

09 Wrap-up

Check understanding, take the homework away, and see where Module 3 goes next.

IN THIS PART

01 Quick-fire check

02 Homework

03 Recap & next module

QUICK-FIRE CHECK

Can you answer these in one sentence?

1. What is the difference between Responsible and Trustworthy AI?
2. Why isn’t fairness always identical treatment?
3. Name three ways bias can enter an AI system.
4. What is a proxy variable?
5. What is a feedback loop?
6. Why can overall accuracy hide unfairness?
7. Transparency vs explainability — what’s the difference?
8. What makes human oversight meaningful?
9. What is automation bias?
10. What is the difference between direct and indirect harm?
11. What is an AI Harm Map?
12. Why should practitioners record unknowns?

HOMEWORK

Complete an AI Harm Map for your Module 1 system

Use the same AI system you described in Module 1. Don’t start again — we’re building your governance evidence module by module.

PART 1

Stakeholders

Users, affected and indirect stakeholders, owner, supplier, risk owner.

PART 2

Potential harms

At least 5: what could happen, to whom, and why.

PART 3

Bias

3 entry points: historical, representation, labelling, measurement, proxy, behaviour, loops.

PART 4

Fairness

3 questions — e.g. does it perform differently across groups?

PART 5

Controls

At least one per major harm.

PART 6

Evidence

Policy, test report, log, training record, dashboard, supplier doc.

PART 7

Unknowns

Write UNKNOWN. An unknown can be a finding.

Worked example in your pack: bank fraud detection, 5 harms mapped.

RECAP

What we learned today

- Responsible AI needs controls and evidence — not statements.
- Trust should be earned through evidence.
- Fairness isn’t always identical treatment.
- Bias can enter at every stage of the lifecycle.
- Neutral-looking variables can act as proxies.
- Feedback loops can reinforce mistakes.
- Overall accuracy can hide unfairness.
- Transparency is about the system; explainability is about an output.
- Accountability must be named.
- Human oversight must be meaningful — watch for automation bias.
- The user isn’t always the person carrying the risk.
- Principle → risk → control → evidence.

CLOSING MESSAGE

Module 1 asked: What is this AI system doing?

Module 2 asks: What could it do to people?

- It can be technically impressive and still cause harm.
- It can be accurate overall and still treat some people unfairly.
- It can have a human in the loop and still suffer automation bias.
- It can have a Responsible AI policy and still have weak controls.

That is how Responsible AI moves from a slogan into governance.

Who is affected?	Can decisions be explained?
What could go wrong?	Who is accountable?
Where could bias enter?	What should a human control?
Could outcomes be unfair?	What controls exist?
Can people understand it?	What evidence do we have?

NEXT MODULE

Module 3 — AI Risk Management and Risk Assessment

Today we identified harm. Next module we turn those harms into structured risks — and build an AI Risk Register.

- risk statements
- causes, events & consequences
- likelihood & impact
- inherent risk
- control effectiveness
- residual risk
- risk appetite & tolerance
- risk ownership & treatment
- AI risk registers
- escalation & monitoring



THE JOURNEY

We keep using HireAI. That is AI Governance, Risk, Compliance and Assurance in practice.

AI GRCA LAB

Thank you

Questions, challenges and real examples welcome.

Facilitated by Collins Obasuyi

PADEMIQ / AI GRCA Lab · pademiq.com / aigrcalab.com