AI Bias Risk Assessment Checklist

(~30 min: 10-15 min for scenario writing + 15-20 min for stage 1)

A Manager's Guide

Complete the scenario section below in groups, then work through the checklist. For each question, note your findings in the Answer / Notes column and assign a risk level. For the online workshop, only section 1 is relevant.

If time is running out, discuss just the highest priority risk question, ***try to cover at least 3 out of the first 6 questions.***

🔴 High risk: act now	🟡 Medium risk: monitor closely	🟢 Low risk: good practice in place

Your AI Scenario

Describe the AI system you are assessing. Be as specific as possible: the more concrete your scenario, the more useful your risk assessment will be.

Project / AI system name	
Description *What does the system do? What is its purpose?*	
Vendor / provider (if any) *Internal build, third-party vendor, open-source model…*	
Vendor claims (if any) *What does the vendor say about accuracy, data, fairness?*	
Current status *Where is the project now? What decision are you being asked to make?*	
Users *Who will operate or interact with the system directly?*	
Affected population *Who will be impacted by the system's outputs, including indirectly?*	
Deployment context *In which country / region? Which regulatory framework applies?*	
Assessed by	
Group	
Date	

AI Scenario: HR screening assistant

Description: Your company, a mid-size European insurance company with about 5,000 employees, is evaluating a vendor-provided GenAI tool that reads incoming CVs and résumés, summarizes them into structured profiles, and produces a "fit score" from 1 to 100 for each candidate relative to the job description. Hiring managers would see the summary and score, then decide who to interview.

Vendor claims: The vendor says the tool was trained on "millions of successful hire profiles across industries." The vendor provides an API, but you do not have access to the training data or the model internals.

Current status: Procurement has identified the vendor. IT has completed a security review. The Head of HR is enthusiastic and wants to pilot the tool for the summer internship hiring cycle, where more than 200 applications are expected for 15 positions. Your team has been asked to approve the pilot.

Users: The HR team, consisting of 3 people, and 8 hiring managers across departments.

Affected population: Job applicants. The summer internship program has historically attracted candidates aged 18 to 25 from across Europe, with diverse educational and socioeconomic backgrounds.

#	Question	Source	Answer / Notes	Risk
STAGE 1: SCOPE & STAKES				
1.1	What decision does this AI influence or automate? Is it truly automated, or human-assisted?	NIST MAP 1.1, 1.2	Influences which candidates get interviews. Effectively a filtering/ranking decision: candidates below a certain score may never be seen by a human.	🔴
1.2	Who is affected? List all stakeholder groups, including indirect ones (customers, employees, third parties).	NIST MAP 5.1, 5.2	**Direct:** Job applicants (screened in or out). **Indirect:** Hiring managers (anchored by scores), HR team (workload shaped by tool), the company (workforce composition over time), teams receiving interns (diversity of perspectives).	🟡
1.3	What happens to a person if the AI gets it wrong? How severe and reversible is the harm?	NIST MAP 5.1; EU AI Act Art. 6	A qualified candidate is rejected without ever being seen by a human. This is **allocative harm:** denial of an opportunity (employment). For interns from disadvantaged backgrounds, this may be their only entry point into the industry. Severity: **HIGH.**	🔴
1.4	Does this qualify as HIGH-RISK under EU AI Act Annex III? (employment, education, credit, biometrics, public services, law enforcement, migration)	EU AI Act Art. 6–7, Annex III	**YES.** EU AI Act Annex III, Point 4(a) explicitly lists AI systems for recruitment or selection of natural persons - placing job ads, filtering applications, evaluating candidates. This is unambiguously high-risk. Full compliance with Articles 9-15 is mandatory. Local employment discrimination law also applies.	🔴
1.5	Was there a formal, documented decision that this should be automated; or did it just happen?	NIST GOVERN 1.2	No. The decision to pilot was driven by HR's workload concern ("too many applications to read"). No one formally assessed whether automation is appropriate for this decision, or what the alternatives are (e.g., structured random sampling, distributing review workload).	🔴
1.6	Have impacted communities or representatives (works council, ERGs, diversity officers) been consulted in problem framing?	NIST MAP 5.2; EU AI Act Art. 9	No. No applicants, employee resource groups, works council, or diversity officers were consulted. The works council may actually have co-determination rights here (depending on jurisdiction).	🔴
STAGE 2: DATA				
2.1	Where does the training / input data come from? Is provenance documented?	EU AI Act Art. 10(2)(b); Gebru et al. (2021)	Vendor says "millions of successful hire profiles across industries." **We have no documentation of what this means.** We don't know: which industries, which countries, which time period, what "successful" means (hired? retained? promoted?), or whether it includes protected characteristics.	🔴
2.2	Who is represented in the data? Who is underrepresented or entirely missing?	Suresh & Guttag - representation bias	**We don't know.** If the training data is predominantly from the US or UK tech industry (common for vendor tools), it may not represent European candidates, non-English-speaking candidates, candidates from non-traditional educational paths, career changers, or people with employment gaps (caregivers, people with disabilities, refugees).	🔴
2.3	Does the data contain proxy variables that correlate with protected characteristics (e.g. postcode → ethnicity, name → gender)?	EU AI Act Art. 10(2)(f); Suresh & Guttag - measurement bias	Almost certainly. CVs contain: university names (→ socioeconomic status, geography, ethnicity), language proficiency levels (→ national origin), extracurricular activities (→ gender, culture, class), employment gaps (→ gender/caregiving, disability), graduation	🔴

#	Question	Source	Answer / Notes	Risk
			dates (→ age). The "fit score" likely encodes these proxies.	
2.4	Is there a valid legal basis for using this data? Have GDPR Art. 22 obligations (automated decision-making) been assessed?	GDPR Art. 6, 9, 22; EU AI Act Art. 10(2)(e)	**GDPR Art. 22 is directly relevant:** applicants have the right not to be subject to purely automated decisions with legal or significant effects. If the score effectively determines who gets interviewed, we may be in violation unless we ensure meaningful human oversight (not rubber-stamping). We need to inform applicants that AI is used. **Currently neither is in place.**	🔴
2.5	Has the vendor provided a datasheet or data card documenting the dataset?	Gebru et al. (2021) - Datasheets for Datasets	**No.** The vendor has not provided a datasheet. We have only marketing materials. We have not asked for one.	🔴
2.6	If using a third-party / vendor model: have you assessed their data documentation and bias testing? Do they accept EU AI Act obligations as provider?	NIST GOVERN 6.1; EU AI Act Art. 25 (deployer liability)	**No.** Procurement did a security review and a price negotiation. No one asked the vendor for bias testing results, fairness evaluations, or data documentation. The vendor's website mentions "AI ethics" but provides no specifics.	🔴
2.7	For non-EU vendors: have you assessed whether they meet EU AI Act obligations and whether your contract reflects deployer responsibilities?	EU AI Act Art. 25; GDPR Art. 28	You inherit full liability for a system designed without EU legal requirements in mind.	🔴
STAGE 3: BUILD & EVALUATION				
3.1	What metric defines 'success' for this system? Who chose it? Is it disaggregated by demographic group?	NIST MEASURE 2.6; Suresh & Guttag - aggregation bias	The vendor reports "92% accuracy in predicting successful hires." **We don't know:** How is "successful hire" defined? Accuracy compared to what baseline? Is this overall or disaggregated? 92% overall accuracy can mask dramatically different performance across groups. **No one on our side has defined what success means for us.**	🔴
3.2	Has performance been evaluated disaggregated by relevant subgroups (gender, age, ethnicity, language, socioeconomic background)?	NIST MEASURE 2.6; EU AI Act Art. 9(7)	**We don't know and we haven't asked.** We need to see performance broken down by at minimum: gender, age group, nationality/ethnicity, educational background, language of CV. The vendor has not volunteered this information.	🔴
3.3	Which definition of fairness applies here, and why? (Equal accuracy? Equal false positive rate? Equal opportunity?)	Chouldechova (2017); Kleinberg et al. (2016); NIST MAP 2.3	**We haven't discussed this at all.** For hiring, the relevant question is: are qualified candidates from different demographic groups equally likely to be scored above the interview threshold? (Equal opportunity.) We haven't set a fairness definition, let alone measured against one.	🔴
3.4	Has the vendor provided a model card?	Mitchell et al. (2019) - Model Cards	**No.** The vendor has not provided one. We have not requested one.	🔴
3.5	Is the evaluation team diverse? Is there psychological safety to flag concerns without career risk?	NIST GOVERN 3.2; Raji et al. (2020)	The evaluation team so far is: Procurement (focused on cost), IT (focused on security), Head of HR (enthusiastic champion). **No one with fairness expertise, no legal review of discrimination risk, no D&I input, no works council involvement.** The Head of HR's enthusiasm may create pressure to approve.	🔴
3.6	For GenAI: has the system been tested for stereotyped outputs, quality differences across languages / demographic groups, and hallucinations?	NIST AI 600-1 (2024)	The tool summarizes CVs using GenAI. **We have not tested whether summaries systematically differ in tone, length, or emphasis based on candidate demographics.** Does it summarize a woman's CV differently than a man's? Does it handle CVs in French, German, or Polish as well as English? Does it hallucinate qualifications? **No testing has been done.**	🔴

#	Question	Source	Answer / Notes	Risk
STAGE 4: LAUNCH GATES				
4.1	Is there a formal go / no-go decision point with explicit fairness criteria before launch?	Raji et al. (2020); EU AI Act Art. 9	**No formal gate.** The current plan is: procurement signs the contract, IT configures the API, HR starts using it. There is no fairness review checkpoint.	🔴
4.2	Who has authority to delay or stop a launch on fairness grounds? Is this person independent from the project champion?	NIST GOVERN 1.4; Raji et al. (2020)	**No one is designated.** The Head of HR is both the champion and the decision-maker. There is no independent check.	🔴
4.3	Are affected users informed that AI is used, how it works, and what data is processed? (Transparency obligation)	EU AI Act Art. 13, 52; GDPR Art. 13–14	**No.** There is currently no plan to inform applicants that AI is used in screening, what it does, or how to contest. This likely violates both the EU AI Act (Art. 13, 52) and GDPR (Art. 13–14, 22).	🔴
4.4	Is there a redress mechanism? Can affected individuals request human review or challenge an AI-driven decision?	EU AI Act Art. 14(4), Art. 86; GDPR Art. 22	**No.** If a candidate is scored low and screened out, there is no process for them to request human review, understand why, or challenge the decision.	🔴
4.5	Has a Responsible AI Impact Assessment been completed?	Microsoft RAI Standard; Raji et al. (2020); EU AI Act Art. 9	**No.** No impact assessment of any kind has been done beyond IT security.	🔴
4.6	Is the right to human review guaranteed and operationalised - not just stated on paper?	GDPR Art. 22; EU AI Act Art. 86	Human oversight exists nominally but not in practice. Automation bias means humans rubber-stamp AI decisions.	🔴
STAGE 5: MONITORING & RESPONSE				
5.1	Is there a monitoring plan with specific metrics, alert thresholds, and a named owner?	NIST MANAGE 1.3, 2.2; EU AI Act Art. 72	**No.** The plan is to "see how it goes during the pilot." No specific metrics, no thresholds, no dashboard.	🔴
5.2	Is performance monitored disaggregated by subgroups on an ongoing basis?	NIST MEASURE 2.6; EU AI Act Art. 9(2)(a)	**No.** Even if we tracked outcomes, we haven't planned to break them down by demographic group.	🔴
5.3	Is there a user feedback channel for fairness concerns - for both affected individuals and internal users?	EU AI Act Art. 14, 86; NIST MANAGE 4.1	**No.** Neither applicants nor hiring managers have a way to flag "this score seems wrong/biased."	🔴
5.4	Is there an incident response runbook? Who can pause the system, and what is the remediation process?	NIST MANAGE 4.1, 4.2; EU AI Act Art. 73	**No.** If a hiring manager notices the tool is systematically scoring certain groups lower, there is no defined escalation path or response procedure.	🔴
5.5	Can the system be paused or rolled back quickly? Is that authority clearly delegated?	NIST MANAGE 4.2; EU AI Act Art. 14(4)(d)	**Probably yes** (it's an API we can turn off), but **no one has been designated with the authority to do so** and there's no documented procedure.	🟡
5.6	Is there a scheduled review cycle (not just 'when something goes wrong')?	NIST MANAGE 3.1; EU AI Act Art. 9(2)(a)	**No.** The pilot has no defined review point. No one has said "after X applications, we stop and evaluate."	🔴

Sources: NIST AI RMF 1.0 (2023) · EU AI Act (2024) · GDPR (2018) · Gebru et al. (2021) Datasheets for Datasets · Mitchell et al. (2019) Model Cards · Raji et al. (2020) · Chouldechova (2017) · Kleinberg et al. (2016) · Suresh & Guttag (2021) · Microsoft RAI Standard · NIST AI 600-1 (2024)