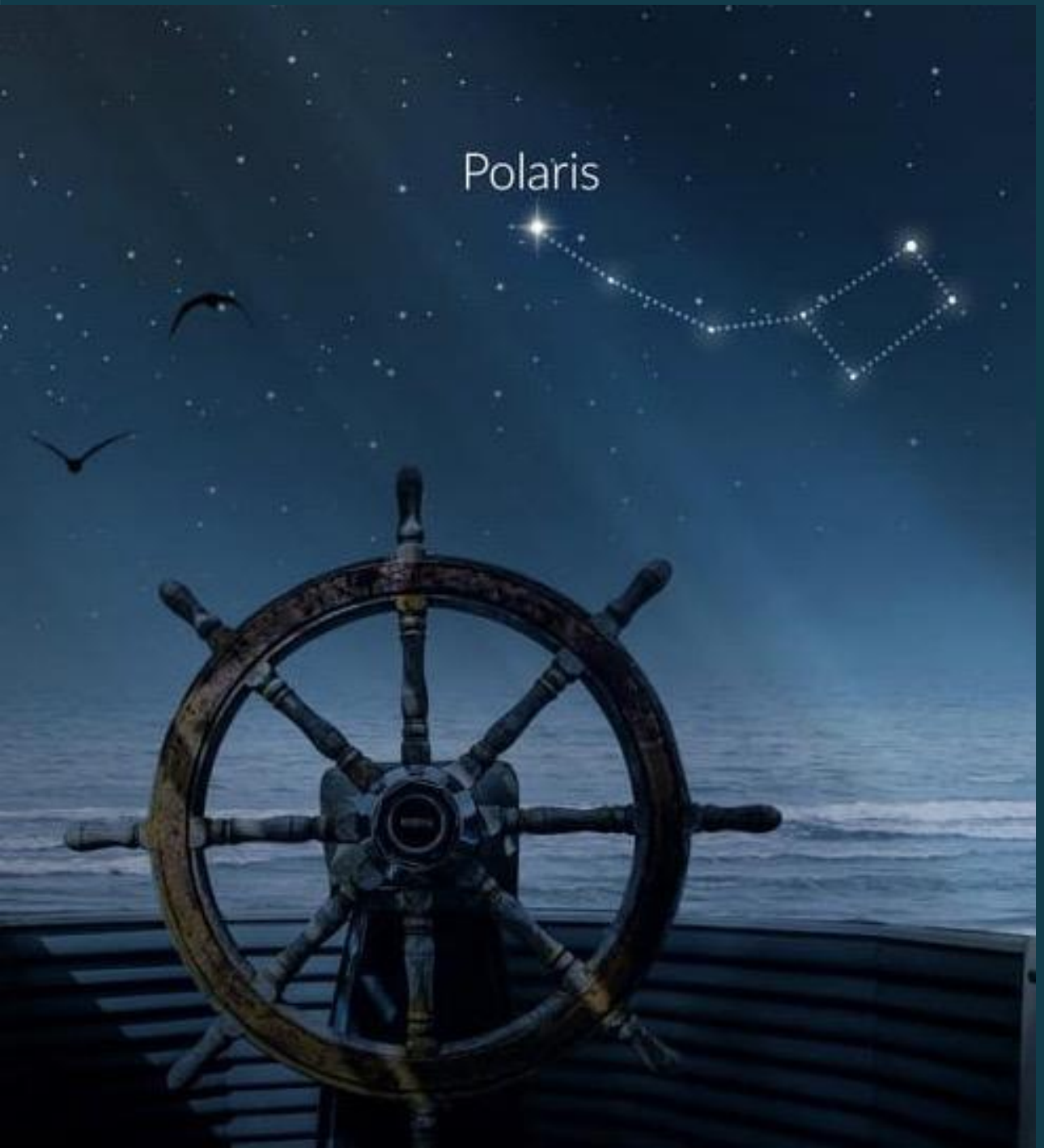
Ethical AI in Talent Assessment

Following True North in our Professional Calling

2025 Annual ACSG Conference

Jaco de Jager
Evalex Product Development

evalex talent solutions ESTABLISHED 1980



Introduction

The True North constellation

Polaris
(The North Star)

constant position in the sky → Symbol of stability

Guiding light
for sailors, travellers

… hope, inspiration

… guidance, direction

ETHICAL AI

(Li, 2023; Kumar, 2024)

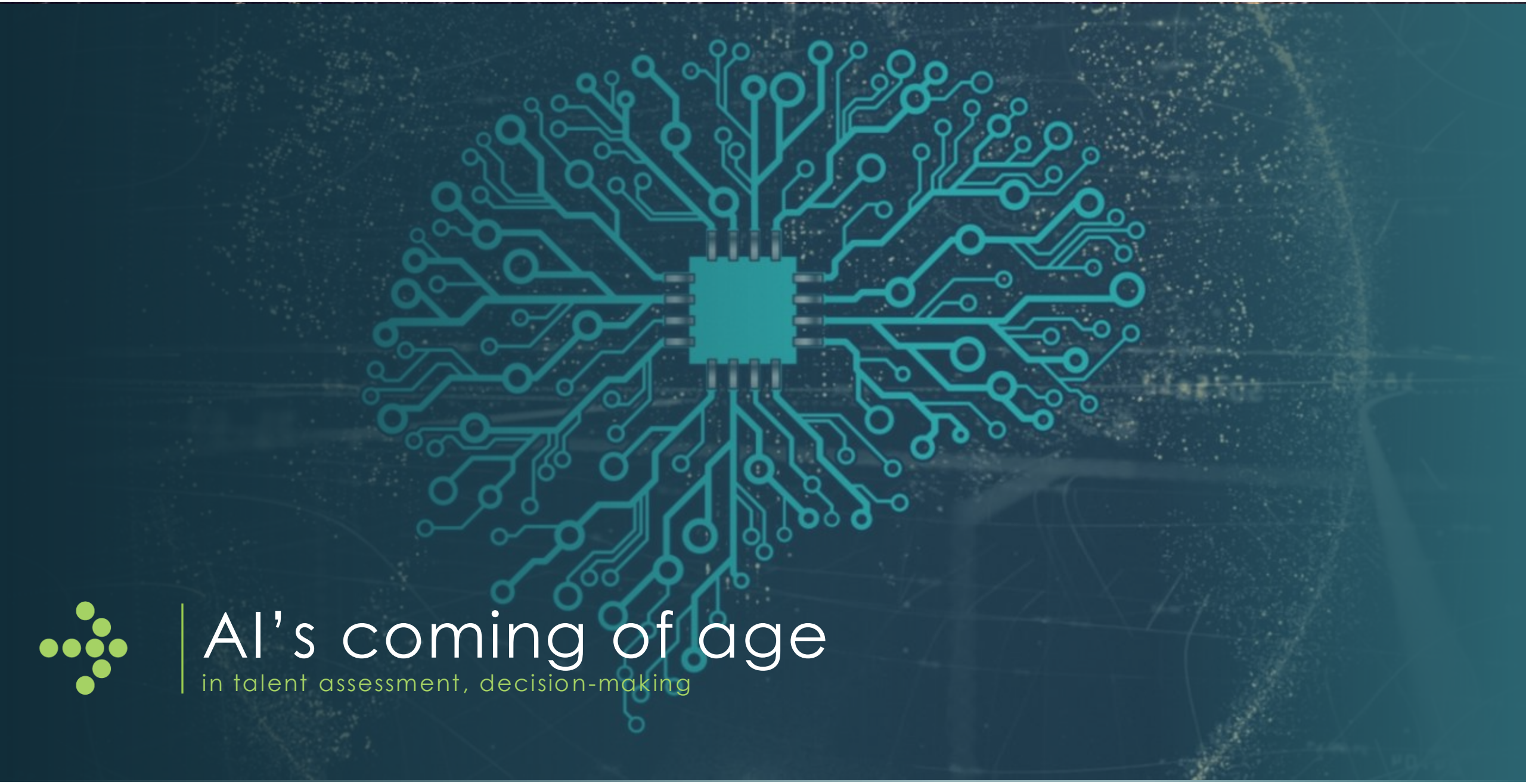
evalex
talent solutions

Agenda

Discussion points

- AI's coming of age in the AC context
  Technology-driven talent decision-making
- Ethical AI – a framework for analysis
  Conceptualisation, key considerations
- False constellations
  Sources, manifestations of AI misconduct
- Following the North Star
  Towards fair, equitable talent assessment
- Conclusions …
  keeping a bearing on True North

evalex
talent solutions



AI's coming of age

in talent assessment, decision-making

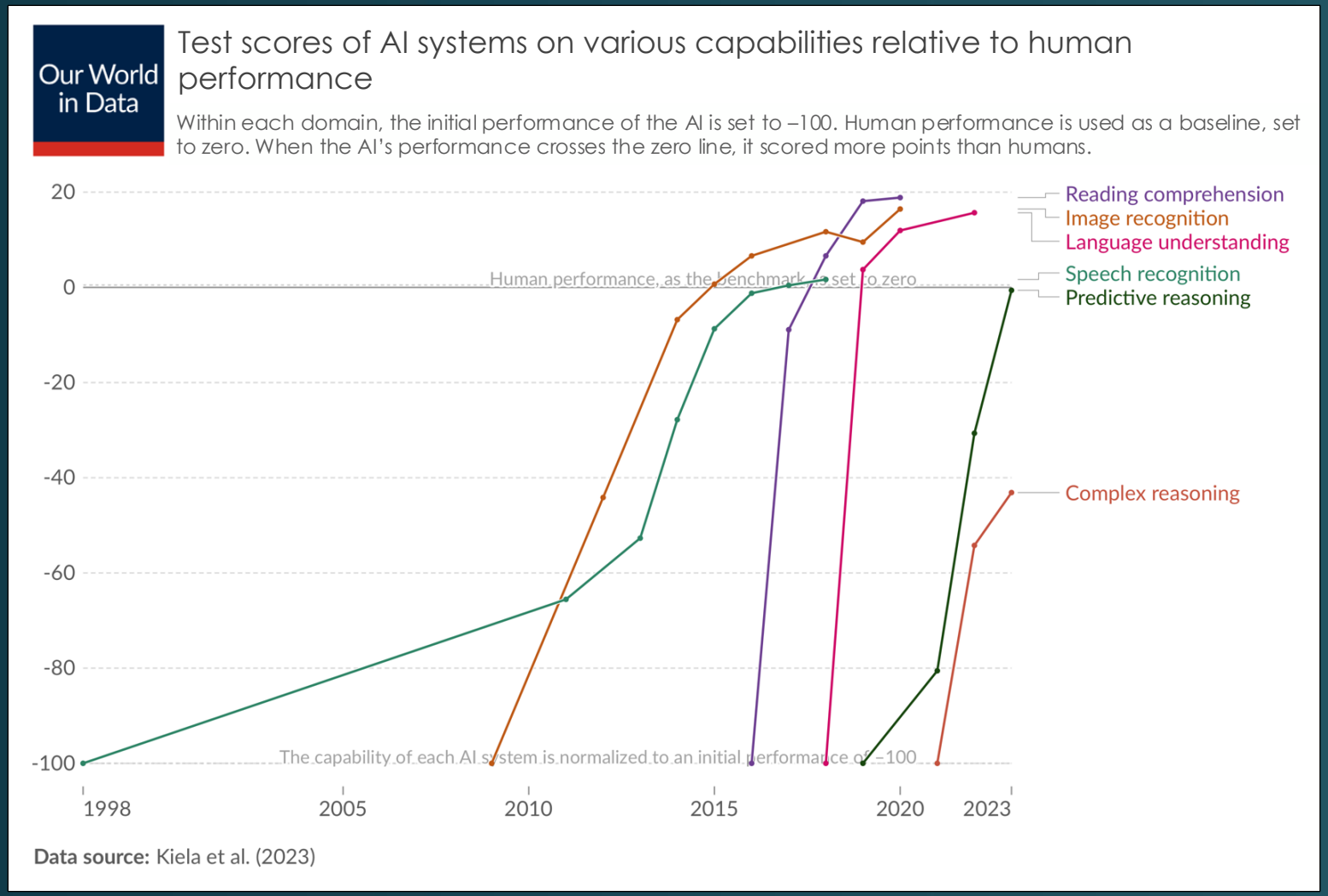
The greatest technological achievement in I–O psychology over the past 100 years?

The development of decision aids that reduce error in predicting employee performance

(Highhouse, 2008)

The evolution of AI

Accelerated development of AI capabilities



AI overtaking average human performance in:

- Image recognition
- Speech recognition
- Reading comprehension
- Language understanding
- Predictive reasoning

(Kiela et al. 2023; Ritchie & Roser, 2024)

The evolution of AI

Accelerated impact on talent decision-making

LEVEL 1:

Automation:
Not truly AI … automating assessment workflow / processes (setup, administration, scoring, reporting)

LEVEL 2:

Job Match/Fit:
Rule-based matching of assessment results against role / level / other criteria; ability to search for capabilities, match people to roles

LEVEL 3:

Algorithmic Interpretation:
Creation of algorithms, rules, systems to interpret reports / results, predict outcomes; multiple layers of interpretation

LEVEL 4:

AI Assessment Products:
Assessment products that are entirely AI-based: assessment, interpretation, reporting; driven by ANNs

LEVEL 5:

True AI - Replacing the psychologist/practitioner:
Using supervised, unsupervised ML to mimic and replace the psychologist in talent processes

Data processing – rule-based, predefined algorithms

Leveraging AI in user-focused scenarios … integration of LLMs

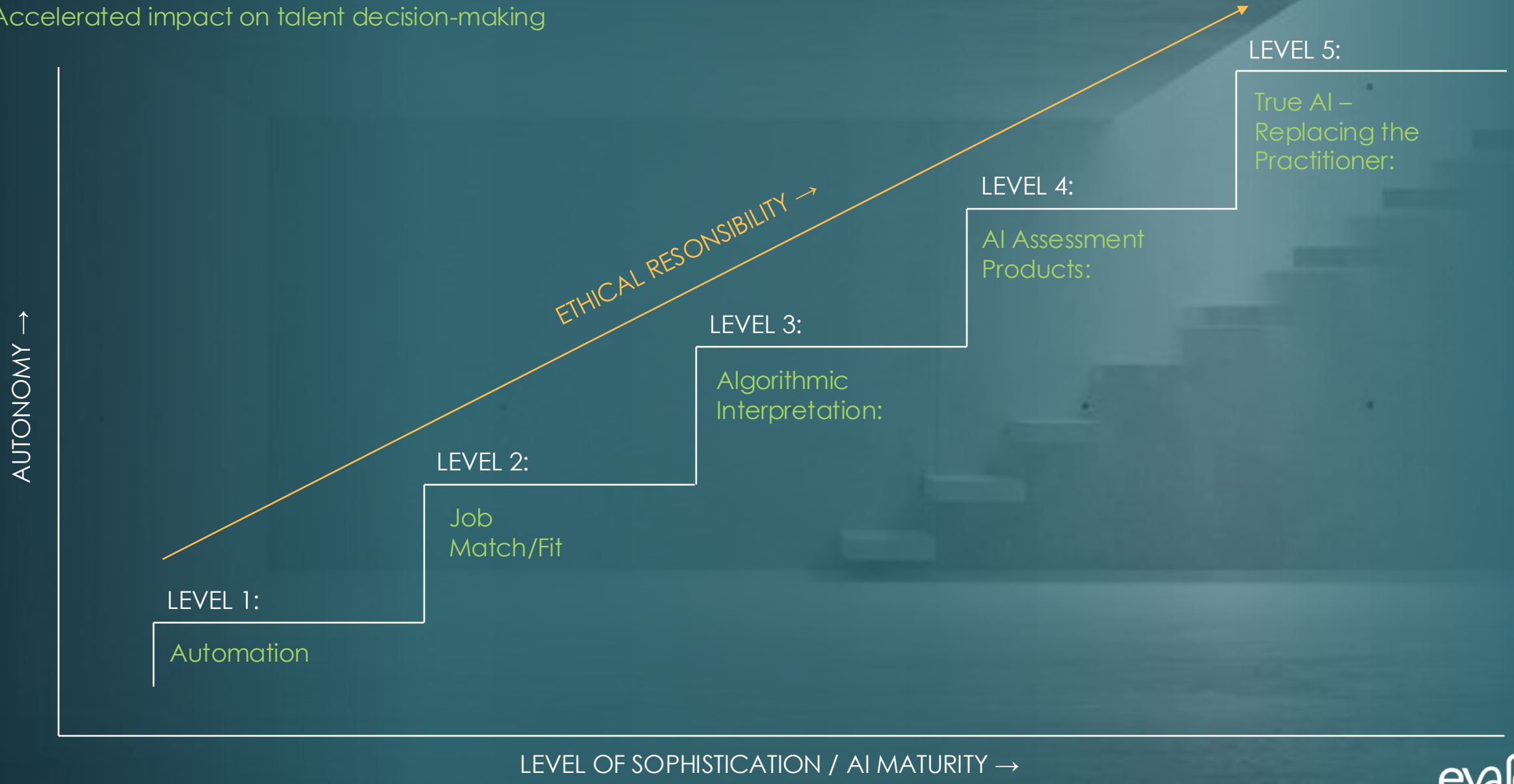
AI designing/optimising/managing other AI … self-modification

(Martín et al., 2024; Mehr, 2023; RECODE, 2021; Shukla et al., 2024)

evalex talent solutions

The evolution of AI

Accelerated impact on talent decision-making

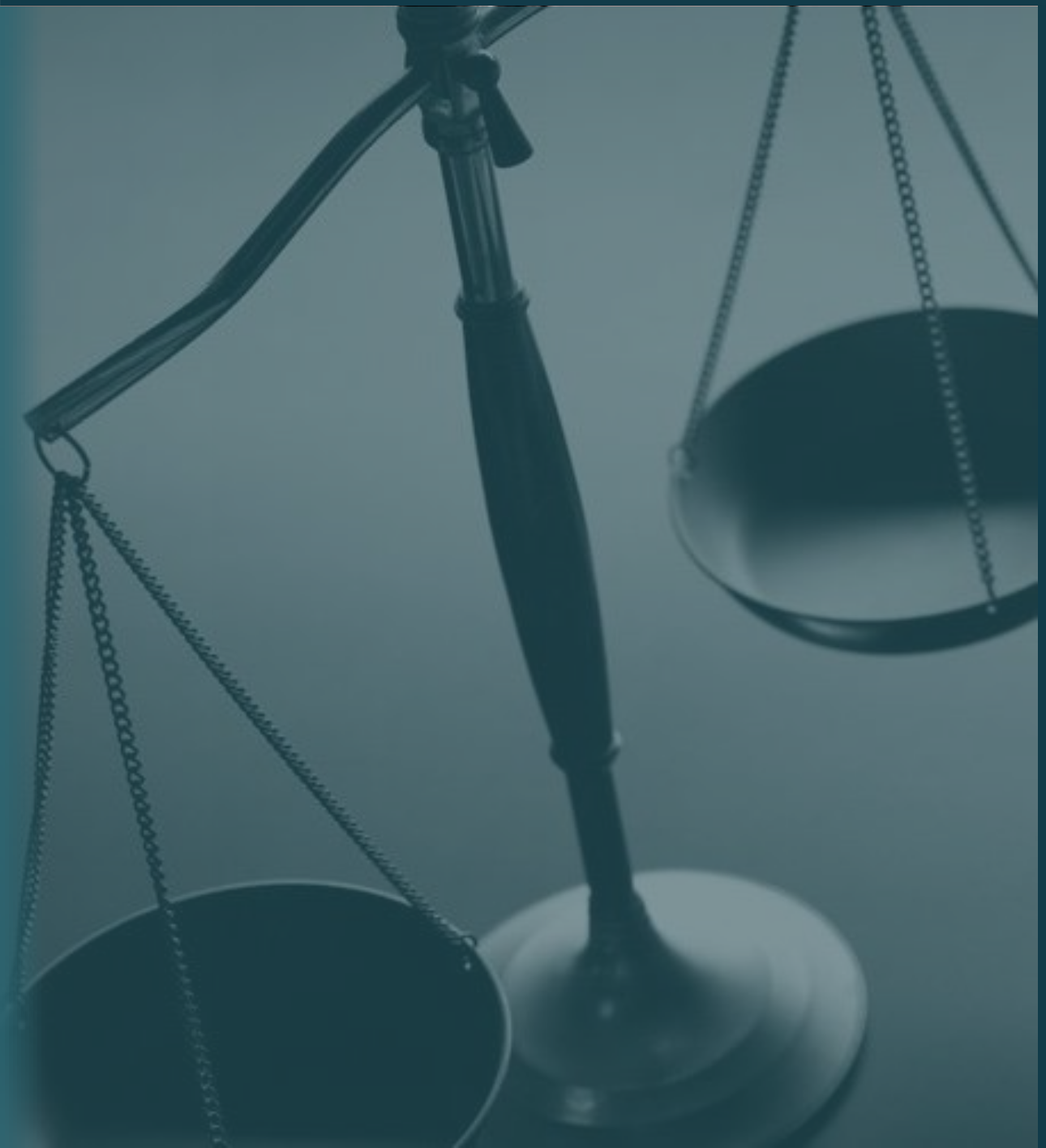


(Shukla et al., 2024; McGraw, 2024)

evalex talent solutions

Ethical AI

A framework for analysis



The journey toward ethical AI in talent management begins with a deep understanding of its core principles.

(Kobayashi, 2024)

Consistency, accuracy

Efficiency gains

Freedom of bias

Fairness

Equity

Accountability

Human oversight, autonomy

Transparency, explainability

Beneficence, non-maleficence

Fidelity, responsibility

Justice, fairness

Respect rights, dignity

Autonomy, informed consent

Do no harm

Serve the greater good

(Hunkenschroer & Luetge, 2022; UNESCO, 2021; Tippins et al., 2021; APA, 2017; Lowman, 2006)

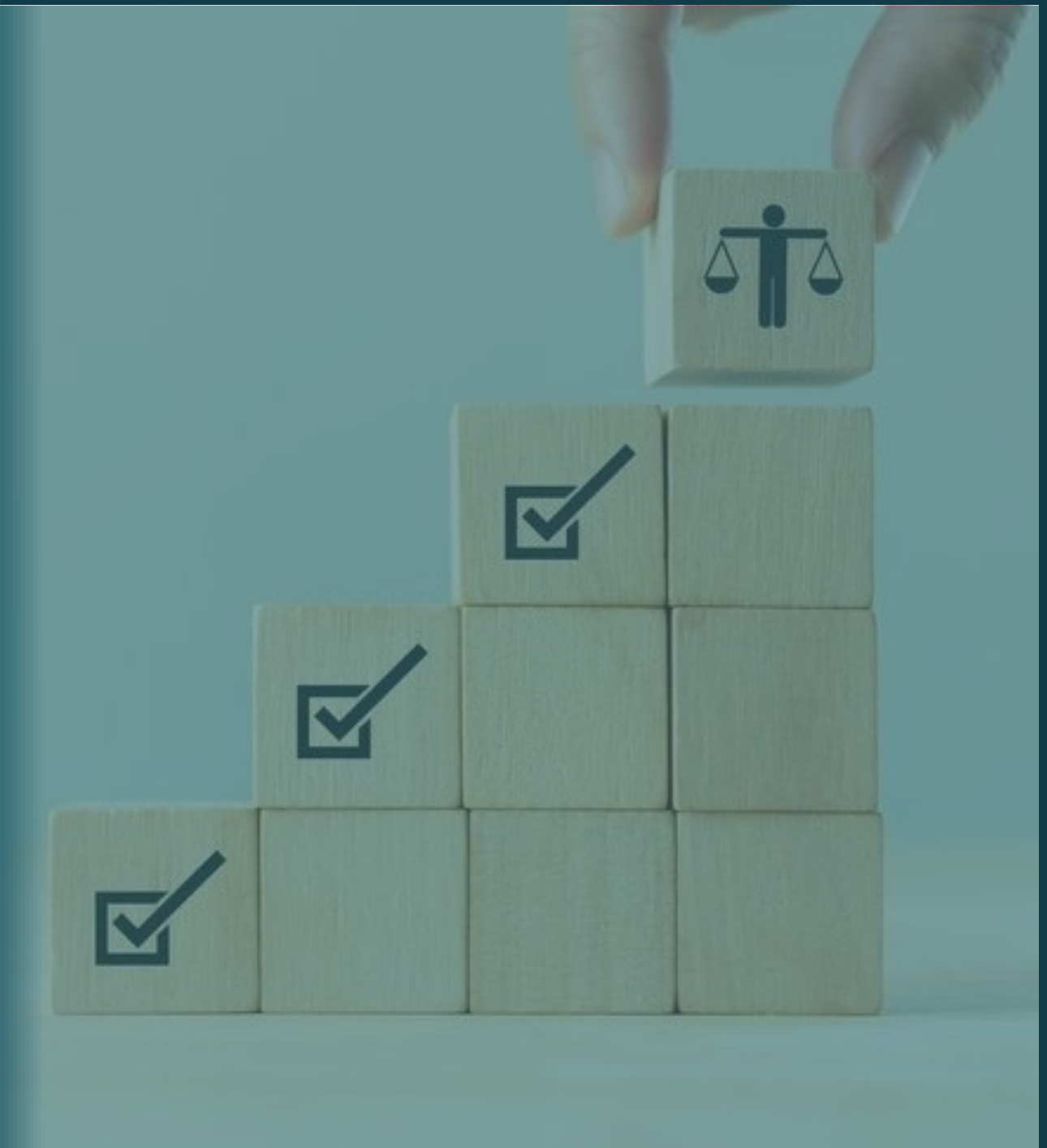
Dimensions of ethical AI

Pointers towards ethical True North

Dimension	Description	Intended Outcomes
Reliability	Consistency, accuracy … algorithmic robustness, dependability	Accurate recommendations, decisions → trust in the results
Fairness	Equitable, bias/discrimination-free … aligned to societal standards	Unbiased recommendations, decisions → equitable opportunities
Privacy	Confidentiality, data protection/control	Protected candidate data → informed consent, anonymity
Transparency	Openness about how AI systems are deployed, monitored, managed	Accessible assessment criteria → defensible assessment process
Explainability	Rationale behind AI decisions … transparent, understandable	Contextualised recommendations, decisions → trust in process
Security, safety	Safeguards against threats, misuse … boundaries, ethical guidelines	Minimised risk from cyber threats, misuse → system reliability
Data governance	Policies, procedures, standards ensuring data quality, security	Governance framework → ethical talent assessment process
Sustainability	Promoting long-term social, economic benefits	Strategic benefits → organisational effectiveness, growth

(AI Index, 2024; Díaz-Rodríguez et al., 2023, UNESCO, 2022; Google, 2018)

evalex talent solutions



Cornerstones

Ethical AI in assessment

Transcends mere compliance

Fostering a culture of trust, fairness

Ensuring a transparent, unbiased process, respectful of individual privacy

(Kobayashi, 2017)

Fake constellations

Sources, manifestations of AI misconduct

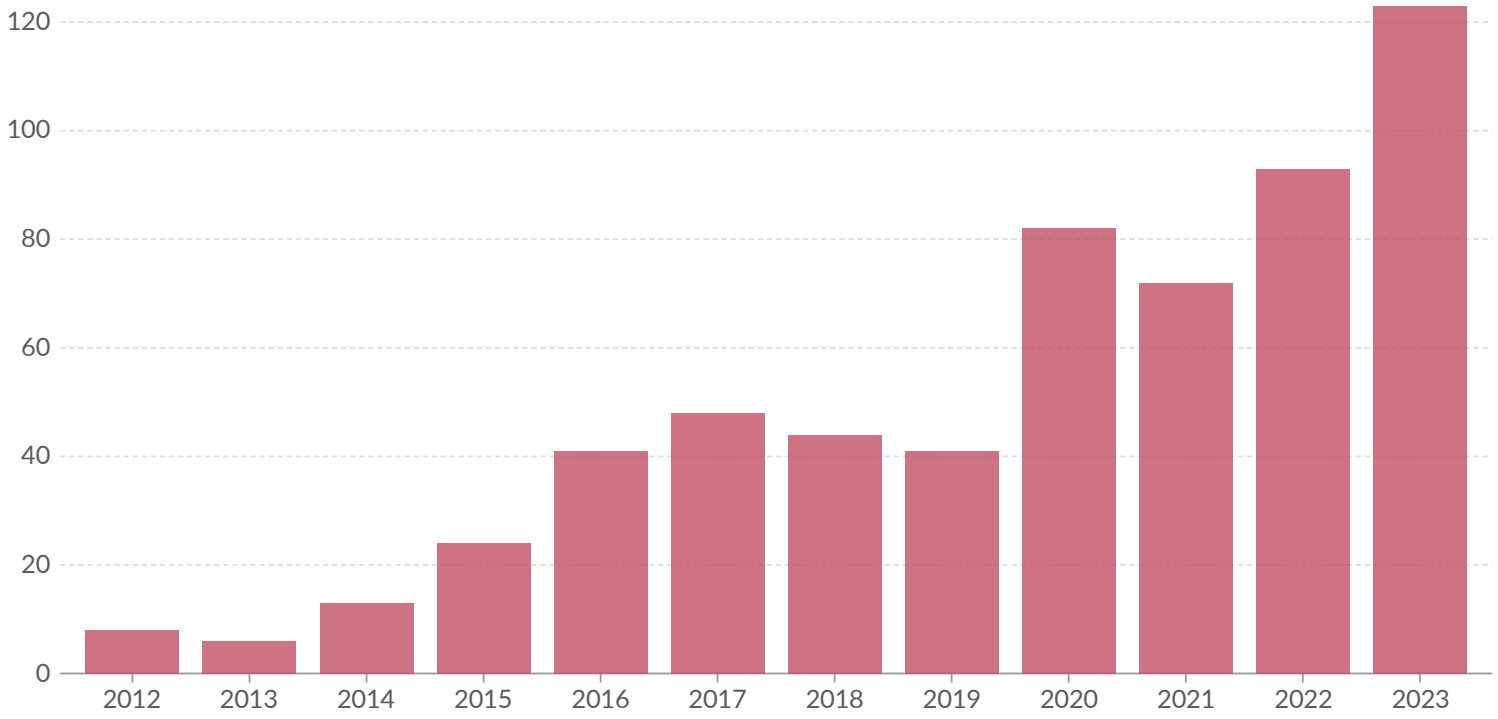
Getting it wrong

The incidence of AI misuse or failure

Global annual number of reported artificial intelligence incidents and controversies

Our World in Data

Notable incidents include a “deepfake” video of Ukrainian President Volodymyr Zelenskyy surrendering, and U.S. prisons using AI to monitor their inmates’ calls.



Data source: AI Incident Database via AI Index (2024) OurWorldinData.org/artificial-intelligence | CC BY

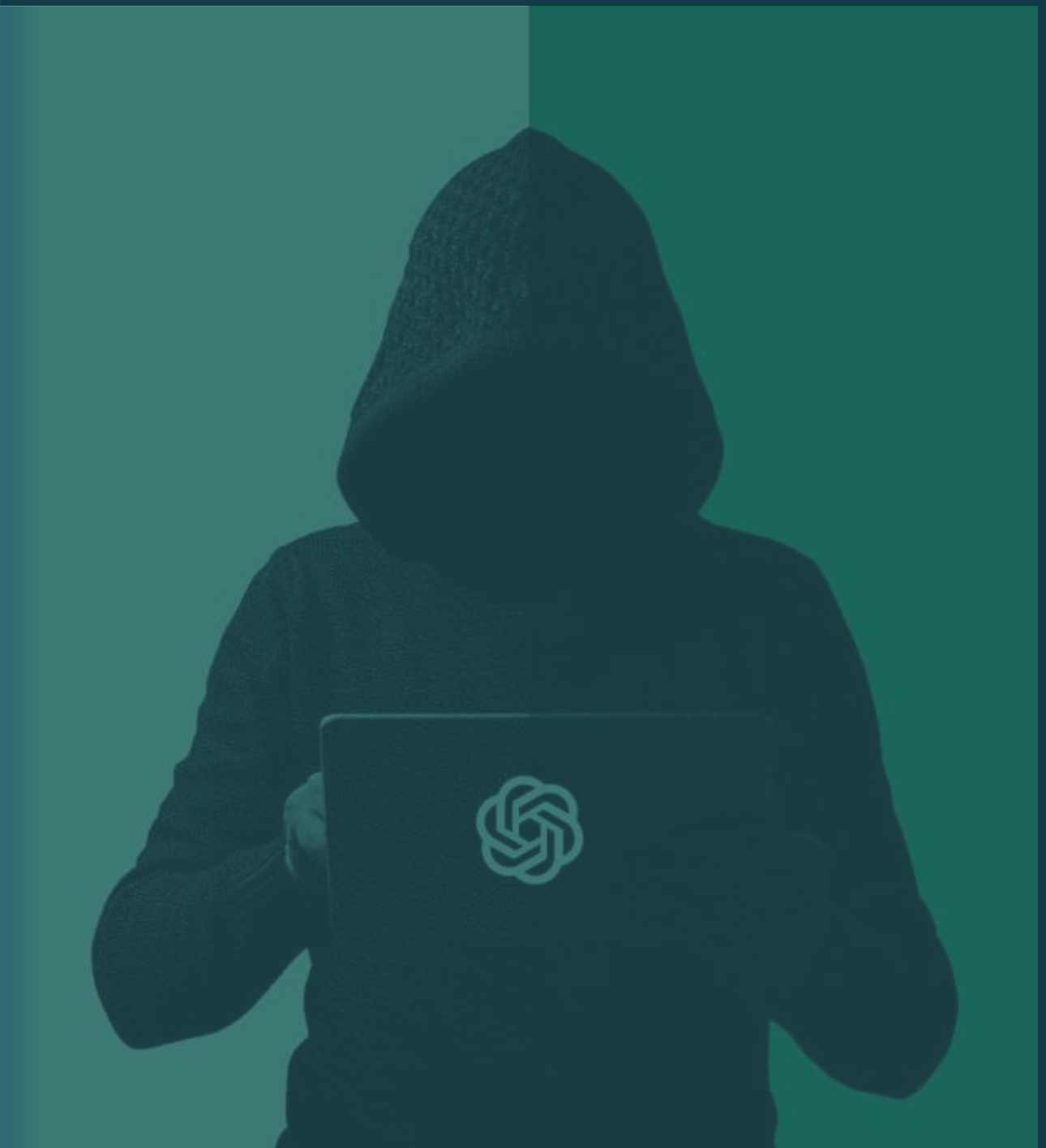
Note: Does not yet include incidents reported in 2022, as incidents must first undergo a vetting process. Reported incidents likely undercount actual incidents, especially in the earliest years of tracking.

AI incidents/controversies on the increase:

- 32.3% increase from 2022 to 2023
- greater real-world integration
- heightened awareness
- previous, current underreporting

(AI Index, 2023; Thanh, 2024)

evalex talent solutions



Disinformation

Generating false/misleading content

Repeated cross-sectional analysis:
vulnerability to manipulation, "hallucination"

Model	Platform	✓ / ✗	Blogs generated	Jailbreaking
GPT-4	via Copilot	●	0/40	0/80
Claude 2	via Poe	●	0/40	0/80
GPT-4	via ChatGPT	●	40/40	(in 38 min)
PaLM 2	via Bard	●	37/40	(in 23 min)
Llama 2	via HuggingChat	●	36/40	(in 51 min)

113 unique blogs, >40 000 words,
purporting false claims

(Menz et al., 2024)

AI bias

Fake constellations

Systematic errors in decision-making

Distorted outputs, harmful outcomes

Inaccurate results
Perpetuate inequalities
Loss of trust
Performance, bottom line
Deteriorating climate, culture
Reputational damage
Legal action
Declining market worth

(Gray et al., 2024; Ferrara, 2024; Highhouse et al,, 2016)

evalex talent solutions

AI bias (in a nutshell)

Unearthing false constellations

1. Algorithmic bias
2. Data bias
3. User bias

Signal Detection Theory

	Actual positive	Actual negative
Predicted positive	TRUE POSITIVE	FALSE POSITIVE
Predicted negative	FALSE NEGATIVE	TRUE NEGATIVE

Accuracy

Proportion of correct classifications
∴ model quality

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad [\approx 1.0]$$

Recall

True Positive Rate (TPR)
∴ probability of detection (sensitivity)

$$\text{TPR} = \frac{TP}{TP + FN} \quad [\approx 1.0]$$

False Positive Rate

Negatives incorrectly seen as positives
∴ probability of false alarm

$$\text{FPR} = \frac{FP}{FP + TN} \quad [\approx 0.0]$$

Precision

Proportion of positives correctly identified
∴ accuracy of positive predictions

$$\text{Precision} = \frac{TP}{TP + FP} \quad [\approx 1.0]$$

F1 score

Harmonic mean of precision, recall
∴ overall performance of AI model

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad [\approx 1.0]$$

(Gray et al., 2024; Ferrara, 2024; Campion & Campion, 2023; Sharma, 2023; Buhl, 2023)

evalex talent solutions

AI's north star

Towards fair, equitable assessment

Ethical AI framework

Opportunities, risks, ambiguities in AI-driven ACs

Ethical opportunities

- Reduction of human bias
- Process consistency
- Timely feedback for applicants
- Efficiency gains for organisations
- Job enhancement for recruiters

Ethical risks

- Introduction of algorithmic bias
- Privacy loss, power asymmetry
- Lack of transparency, explainability
- Obfuscation of accountability
- Potential loss of human oversight

Ethical ambiguities

- Assessment validity, accuracy
- Perceived fairness
- Effect on workforce diversity
- Informed consent
- Use of personal data

Institutional regulation

Organisation standards

Employee/applicant awareness

Technical due diligence

(Hunkenschroer & Luetge, 2022; Tippins et al., 2021)

Ethical AI checklist

Pointers for maintaining a bearing on True North

RELIABILITY:

- Mitigation measures: model errors, low confidence outputs
- Failover plans: system/ model availability
- Model/system evaluation: vulnerabilities, harmful behaviour ('red teaming')
- Safeguards: adversarial attacks
- Confidence scoring: model output reliability
- Sensitivity analysis: range of scenarios, metrics

FAIRNESS:

- Representative data: anticipated user demographics
- Independent oversight: auditable methodology, data sources
- Diverse stakeholders: model development, review
- Subgroup analysis: adverse impact analysis
- Technical bias mitigation: fairness-aware algorithms

TRANSPARENCY:

- Documented dev process: algorithm design choices, data sources, intended use cases, limitations
- User education: intended use cases, model limitations
- Model simplicity: Interpretability vs. performance
- Model explainability tools: saliency maps to elucidate model decisions

SECURITY:

- Cybersecurity hygiene practices: multifactor authentication, access controls, user trg
- 3rd party vetting: validating cybersecurity measures
- AI cybersecurity team: trained for AI-specific cybersecurity
- AI-specific cybersecurity checks: adversarial testing, vulnerability assessments
- Systems integration: monitoring evolving AI-specific cybersecurity risks

DATA governance:

- Compliance checks: relevant laws, regulations; used with consent
- Data validation: completeness, uniqueness, consistency, accuracy
- Contextual relevance: Fairness metrics. disparate impact metrics
- Dataset documentation: traceability throughout AI lifecycle
- MRM frameworks: documenting, remedying deficient datasets

(Gray et al., 2024; Ferrara, 2024; Kobayashi, 2024; Thanh, 2024)

evalex talent solutions

To summarise

Maintaining a bearing on True North

Strategies to reduce bias in AI-driven talent assessment systems:

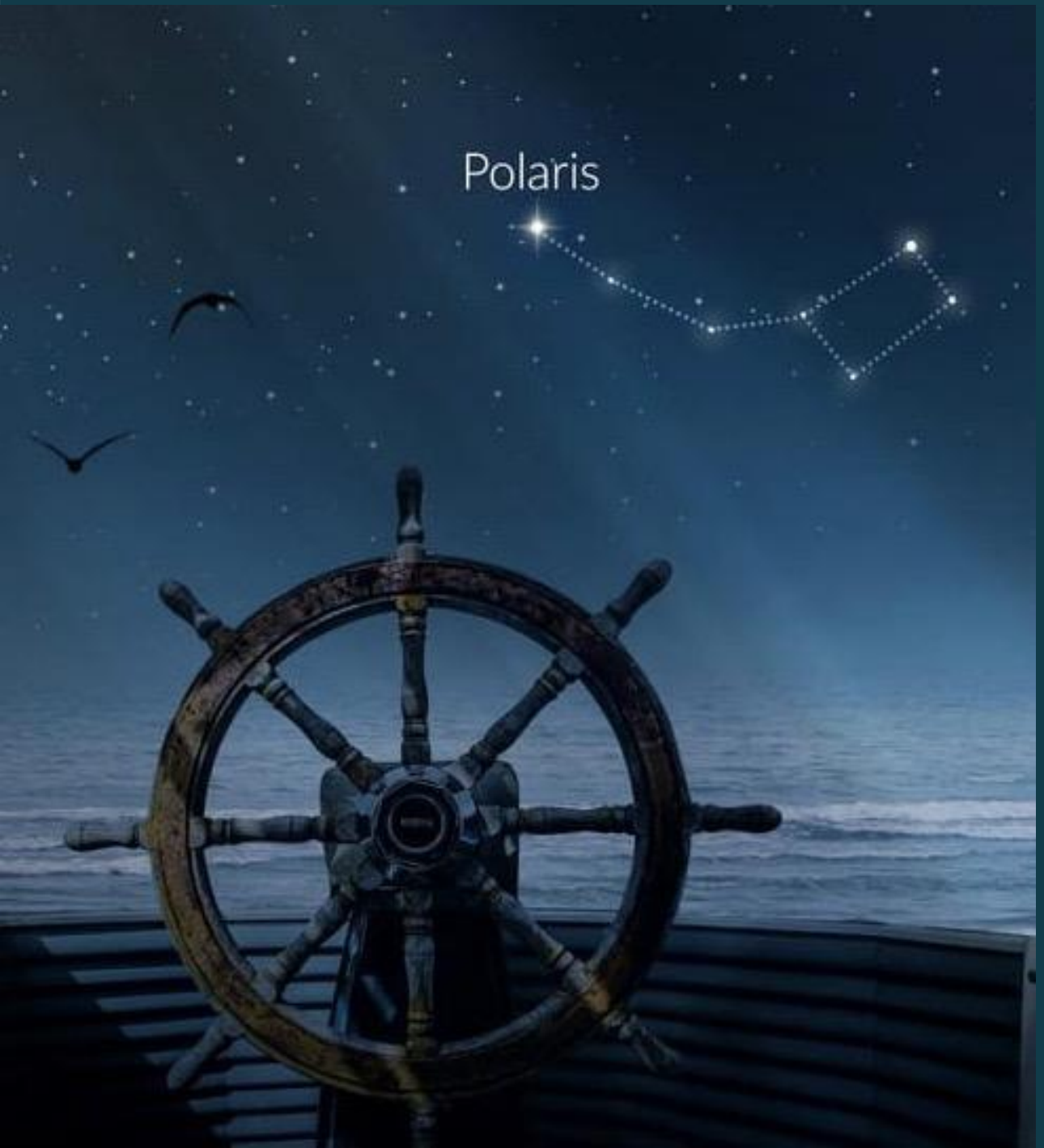
- Diverse data sets:
  Diverse, representative datasets for training AI models to avoid discriminatory patterns
- Algorithmic audits:
  Regular audits of to identify, address potential biases; independent reviews to ensure impartiality
- Inclusive design:
  Diverse inputs in design, development to bring a range of perspectives, mitigate unconscious biases
- Continuous monitoring:
  Feedback mechanisms to identify, correct biases as AI systems are deployed

(Gray et al., 2024; Ferrara, 2024; Kobayashi, 2024)

evalex talent solutions

Conclusion

Keeping a bearing on True North



Ethical AI in the assessment domain is not merely a compliance/regulatory necessity but a strategic imperative:

… assessment systems reliable; perform optimally; adhere to professional standards,

… safeguarding reputation, fostering user trust

… preventing costly errors; reduce the risk of deploying flawed assessment solutions that could lead to operational, legal challenges

evalex
talent solutions

The end

Thank you

Jaco de Jager
Evalex Talent Solutions
jaco@evalex.com

References

Ethical AI in Talent Assessment

AI Index. (2024). Artificial Intelligence Index Report 2024. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/ai-index/2024-ai-index-report

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct (2002, amended effective June 1, 2010, and January 1, 2017). https://www.apa.org/ethics/code/

Buhl, N. (2023, July 18). F1 score in machine learning. Encord. https://encord.com/blog/f1-score-in-machine-learning/#:~:text=making%20more%20accurately.-,Interpreting%20the%20F1%20Score,model%20accurately%20predicted%20each%20label.&text=What%20is%20a%20good%20F1,a%20more%20suitable%20evaluation%20metric.

Campion, M. A., & Campion, E. D. (2023). Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. Personnel Psychology, 76(4), 993–1009. https://doi.org/10.1111/peps.12621

Diaz-Rodríguez, N., Ser, J. D., Coeckelbergh, M., Prado, M. L., Herrera-Viedma, E. E., & Herrera, F. (2023). Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion 99, 101896. https://doi.org/10.48550/arXiv.2305.02231

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1), 3. https://doi.org/10.3390/sci6010003

Google (2024). End-to-end responsibility: A lifecycle approach to AI. https://ai.google/static/documents/ai-responsibility-2024-update.pdf

Gray, M., Samala, R., Liu, Q., Skiles, D., Xu, J., Tong, W., & Wu, L. (2024). Measurement and mitigation of bias in artificial intelligence: A narrative literature review for regulatory science. Clinical Pharmacology & Therapeutics, 115(4), 687–697. https://doi.org/10.1002/cpt.3117

evalex talent solutions ESTABLISHED 1980

References

Ethical AI in Talent Assessment

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. Industrial and Organizational Psychology, 1(3), 333-342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Highhouse, S., Doverspike, D., & Guion, R. M. (2016). Analyzing bias and ensuring fairness: Unfair discrimination, item and test bias, and reducing adverse impact. In: Essentials of personnel assessment and selection (2nd ed.). Routledge/Taylor & Francis Group.

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of ai-enabled recruiting and selection: A review and research agenda. Journal of Business Ethics. Advance online publication. https://doi.org/10.1007/s10551-022-05049-6

Kiela, D., Thrush, T., Ethayarajh, K., & Singh, A. (2023). Plotting progress in AI. Contextual AI Blog. Available at: https://contextual.ai/blog/plotting-progress (Accessed: 02 April 2024).

Kobayashi, T. (2024, February 22). Ethical AI in talent management: Navigating the new frontier. Medium. https://medium.com/@takatsugu/ethical-ai-in-talent-management-navigating-the-new-frontier-0549281e728d

Kumar, B. (2024, November 28). Understanding the concept of a "north star": Meaning, uses, and everyday relevance. Medium. https://bk10895.medium.com/understanding-the-concept-of-a-north-star-meaning-uses-and-everyday-relevance-6d424eec66f3#:~:text=Metaphorically%2C%20the%20term%20"North%20Star,things%20seem%20confusing%20or%20overwhelming.

Li, F. F. (2023). The worlds I see: Curiosity, exploration, and discovery at the dawn of AI. First edition. Moment of Lift Books.

Lowman, R. L. (Ed.). (2006). The ethical practice of psychology in organizations (2nd ed.). American Psychological Association. https://doi.org/10.1037/11386-000

evalex talent solutions ESTABLISHED 1980

References

Ethical AI in Talent Assessment

Martín, J. C., Caba, M. A. G., Peña, L. F., & Láinez, S. J. (2024). The rise of large language models: From fundamentals to application. Management Solutions. Retrieved from https://www.managementsolutions.com/sites/default/files/minisite/static/72b0015f-39c9-4a52-ba63-872c115bfbd0/llm/pdf/rise-of-llm.pdf

McGraw, D. K. (2024). Ethical responsibility in the design of artifificial intelligence (AI) systems. International Journal on Responsibility, 7(1), Article 4. https://doi.org/10.62365/2576-0955.1114

Mehr, M. (2023). The evolution of AI: Unveiling the 7 stages from rule-based systems to the enigmatic AI singularity. Medium, July 23. Retrieved from https://maryammehr345.medium.com/the-evolution-of-ai-unveiling-the-7-stages-from-rule-based-systems-to-the-enigmatic-ai-singularity-e0425ae0858c

Menz, B. D., Kuderer, N. M., Bacchi, S., Modi, N. D., Chin-Yee, B., Hu, T., Rickard, C., Haseloff, M., Vitry, A., & McKinnon, R. A. (2024). Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: Repeated cross sectional analysis. BMJ 2024, 384: e078538. https://doi.org10.1136/bmj- 2023-078538

RECODE. (2021). The 7 stages of the future evolution of artificial intelligence. RECODE, February 28. Retrieved from https://recode.net/the-7-stages-of-the-future-evolution-of-artificial-intelligence/

Ritchie, H., & Roser, M. (2024). Test scores of AI systems on various capabilities relative to human performance. Our World in Data. Retrieved from https://ourworldindata.org/grapher/test-scores-ai-capabilities-relative-human-performance

Sharma, N. (2023, June 06). Understanding and applying F1 score: AI evaluation essentials with hands-on coding example. Arize AI, Inc. https://arize.com/blog-course/f1-score/

Shukla, A. K., Terziyan, V., & Tiihonen, T. (2024). AI as a user of AI: Towards responsible autonomy. Heliyon, 10(11), e31397. https://doi.org/10.1016/j.heliyon.2024.e31397

evalex talent solutions ESTABLISHED 1980

References

Ethical AI in Talent Assessment

Thanh, D. B. (2024, December 20). AI model testing: Crafting Reliable AI models for tomorrow. Smartdev. https://smartdev.com/ai-model-testing-guide/

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action. Personnel Assessment and Decisions, 7(2), 1–22. https://doi.org/10.25035/pad.2021.02.001

UNESCO (2021). Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization, 7 Place de Fontenoy, 75352 Paris 07 SP, France. https://unesdoc.unesco.org/ark:/48223/pf0000381137

evalex talent solutions ESTABLISHED 1980