LECTURE 08 | المحاضرة 08

AI Ethics

أخلاقيات الذكاء الاصطناعي

Responsible, Safe, Transparent, and Trustworthy Artificial Intelligence

الذكاء الاصطناعي المسؤول والآمن والشفاف والجدير بالثقة

Prepared by | إعداد: Dr.-Ing. Aouss Gabash

🎯 Learning Objectives

- Understand why AI ethics is an engineering issue.
- Explain fairness, transparency, explainability, and accountability.
- Understand data privacy and cybersecurity risks.
- Recognize bias in data and AI models.
- Understand the importance of human oversight.
- Discuss AI safety in critical power-system infrastructure.

🎯 أهداف التعلم

- فهم لماذا تعد أخلاقيات الذكاء الاصطناعي مسألة هندسية.
- فهم العدالة والشفافية وقابلية التفسير والمساءلة.
- فهم مخاطر خصوصية البيانات والأمن السيبراني.
- التعرف على التحيز في البيانات والنماذج.
- فهم أهمية الإشراف البشري.
- مناقشة سلامة الذكاء الاصطناعي في البنية التحتية الكهربائية الحرجة.

1. Why Do We Need AI Ethics?

AI systems increasingly influence decisions that affect people, organizations, infrastructure, and society.

> A model can be mathematically accurate and still be unsafe, unfair, misleading, or inappropriate for real-world use.

1. لماذا نحتاج إلى أخلاقيات الذكاء الاصطناعي؟

تؤثر أنظمة الذكاء الاصطناعي بصورة متزايدة في قرارات تمس الأفراد والمؤسسات والبنية التحتية والمجتمع.

> قد يكون النموذج دقيقًا رياضيًا، ومع ذلك قد يكون استخدامه غير آمن أو غير عادل أو مضللًا.

2. Responsible AI

Safe
The system should not create unacceptable physical or operational risks.

Fair
Decisions should not systematically disadvantage particular groups.

Transparent
Users should understand how and why AI is being used.

Accountable
Responsibility for AI-supported decisions must remain identifiable.

2. الذكاء الاصطناعي المسؤول

يجب أن يكون النظام المسؤول آمنًا وعادلًا وشفافًا وقابلًا للمساءلة.

3. Data Bias

$$ModelBehavior \approx LearnedPatternsfromTrainingData$$

Biased or incomplete training data can produce biased predictions.

In AI, biased data can produce convincing but systematically incorrect outputs.

3. التحيز في البيانات

قد تؤدي البيانات غير المتوازنة أو غير الممثلة جيدًا للواقع إلى قرارات منحازة.

جودة البيانات جزء أساسي من موثوقية النظام.

4. Dataset Shift

A model trained on normal weekdays may behave poorly during holidays, heat waves, outages, or emergencies.

$$Ptraining(x) \neq Preal - world(x)$$

Good training performance does not guarantee good performance when real-world conditions change.

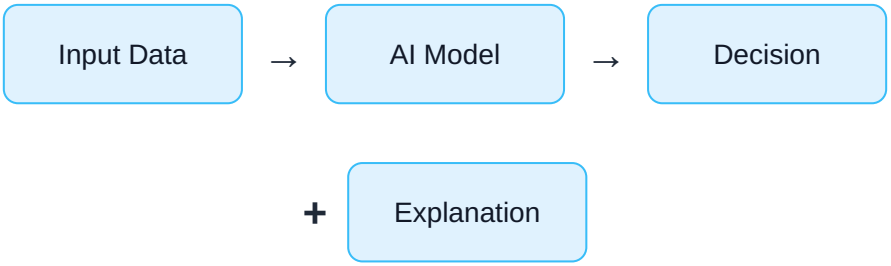
4. تغير توزيع البيانات

قد يختلف سلوك البيانات في العطل أو موجات الحر أو حالات الطوارئ عن بيانات التدريب.

$$Ptraining(x) \neq Preal - world(x)$$

5. Explainability

Explainability means providing understandable reasons for an AI prediction or recommendation.



> In engineering, “the model said so” is often not sufficient.

5. قابلية التفسير

تعني توفير سبب مفهوم للقرار أو التنبؤ الذي قدمه النموذج.

> في التطبيقات الهندسية لا يكفي عادةً القول إن النموذج أعطى هذه النتيجة.

6. Accuracy vs Explainability

Simple Model
Often easier to understand and audit.

Complex Model
May capture more complex relationships but can be harder to explain.

> The best engineering model is not always the model with the smallest prediction error.

6. الدقة مقابل قابلية التفسير

قد يكون النموذج الأكثر تعقيدًا أكثر دقة لكنه أصعب في التفسير.

> في التطبيقات الحرجة يجب موازنة الدقة مع الشفافية والسلامة.

7. Human-in-the-Loop

Measurements → AI Recommendation → Engineer Review →

Action

7. الإنسان ضمن حلقة القرار

يجب أن يبقى مهندس أو خبير مؤهل مشاركًا في القرارات المهمة.

الذكاء الاصطناعي يدعم القرار الهندسي، لكنه لا يلغي مسؤولية المهندس.

8. Automation Bias

Automation bias occurs when people trust an automated recommendation simply because it was produced by a computer system.

High confidence from an AI system is not proof that the result is correct.

8. التحيز نحو الأتمتة

يحدث عندما يثق المستخدم بتوصية النظام تلقائيًا لمجرد أن الحاسوب أنتجها.

ثقة النموذج العالية لا تعني بالضرورة أن النتيجة صحيحة.

9. Privacy

AI systems may process operational, customer, household, and infrastructure data.

- Data minimization
- Purpose limitation
- Secure storage

- Controlled access
- Appropriate anonymization

9. خصوصية البيانات

يجب جمع البيانات الضرورية فقط وتأمينها والتحكم بمن يستطيع الوصول إليها.

10. Cybersecurity

Data Manipulation
An attacker may alter measurements or training data.

Model Manipulation
Model parameters or software may be modified.

Adversarial Inputs
Designed inputs may cause incorrect predictions.

Unauthorized Access
Access may turn an AI tool into an operational risk.

10. الأمن السيبراني

تشمل المخاطر التلاعب بالبيانات أو النموذج أو المدخلات والوصول غير المصرح به.

11. AI in Critical Infrastructure

An incorrect recommendation in a power network can have physical consequences.

Power-system AI therefore requires strict verification, validation, monitoring, and fallback mechanisms.

11. الذكاء الاصطناعي في البنية التحتية الحرجة

قد يؤدي الخطأ في قرار حماية أو تحكم أو تشغيل كهربائي إلى نتائج مادية حقيقية.

12. Safety Constraints

$$Vmin \leq V \leq Vmax$$

$$I \leq Imax$$

$$SOCmin \leq SOC \leq SOCmax$$

AI optimization should operate inside an engineering safety envelope.

12. قيود السلامة

يجب ألا يسمح لنظام الذكاء الاصطناعي بانتهاك الحدود الفيزيائية الآمنة.

يجب أن يعمل الذكاء الاصطناعي داخل غلاف سلامة هندسي محدد.

13. Fail-Safe Design

AI Output → Safety Check → Accept or Reject

Unsafe / Invalid → Fallback Controller

13. التصميم الآمن عند الفشل

يجب تحديد ما يحدث عندما يفشل النموذج أو ينتج قيمة غير منطقية.

يمكن العودة إلى متحكم تقليدي موثوق أو نقطة تشغيل آمنة.

14. Accountability

- Who approved the model?
- Who validated the data?
- Who monitors performance?
- Who can override the system?
- Who investigates failures?

14. المساءلة

يجب تحديد من يعتمد النموذج ويراقبه ويوقفه ويتحمل مسؤولية التحقيق عند الخطأ.

15. Monitoring After Deployment

$$\text{Training} \longrightarrow \text{Validation} \longrightarrow \text{Deployment} \longrightarrow \text{Monitoring} \longrightarrow \text{Revalidation}$$

Model performance can deteriorate as system conditions change.

15. المراقبة بعد التشغيل

لا تنتهي عملية تقييم النموذج بمجرد تشغيله، فقد يتدهور الأداء مع تغير الظروف.

16. Engineering Verification

- Normal and extreme operating conditions
- Missing or corrupted measurements

- Out-of-distribution inputs
- Communication and sensor failures

> Testing only average conditions is not enough for safety-critical engineering.

16. التحقق الهندسي

يجب اختبار النموذج في الحالات الطبيعية والمتطرفة ومع فقدان البيانات وأعطال الحساسات والاتصالات.

17. Example: AI Voltage Controller

$$\Delta QAI = f(V, P, Q, Load, PV)$$

AI Recommendation → Voltage Check → Current Check →

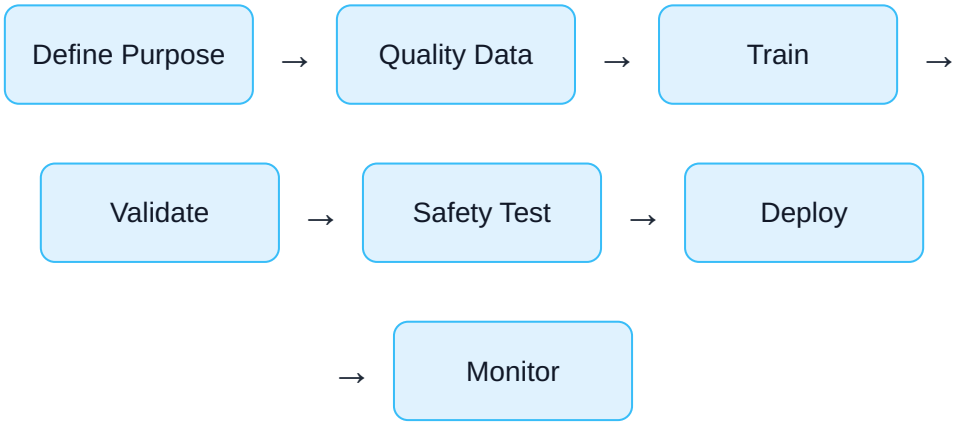
Apply Action

> Intelligence does not replace physical constraints.

17. مثال: متحكم جهد يعتمد على AI

قبل تنفيذ أمر القدرة الردية يجب التأكد من بقاء الجهد والتيار ضمن الحدود المسموحة.

18. Responsible AI Workflow



18. دورة تطوير ذكاء اصطناعي مسؤول

تحديد الهدف ← جمع بيانات جيدة ← التدريب ← التحقق ← اختبار السلامة ← التشغيل ← المراقبة

19. Key Questions Before Deployment

1. Is the model sufficiently accurate?
2. Do we understand its failure modes?
3. Are training data representative?
4. Can the decision be explained?
5. Are safety limits enforced independently?
6. Can a human override the model?
7. Is there a fallback strategy?
8. Is the system cybersecure?
9. How will performance be monitored?

19. أسئلة أساسية قبل التشغيل

1. هل دقة النموذج كافية؟
2. هل نعرف حالات الفشل؟
3. هل بيانات التدريب ممثلة للواقع؟
4. هل يمكن تفسير القرار؟
5. هل قيود السلامة مستقلة؟
6. هل يمكن للإنسان تجاوز القرار؟
7. هل يوجد نظام احتياطي؟
8. هل النظام محمي سيبرانيًا؟
9. كيف ستتم مراقبة الأداء؟

20. Summary

- AI ethics is part of responsible engineering.
- Biased data can produce biased behavior.
- Explainability matters in high-impact applications.
- AI does not eliminate human accountability.
- Privacy and cybersecurity matter throughout the lifecycle.
- Power-system AI must operate inside explicit safety constraints.
- Fail-safe mechanisms and human oversight are essential.
- Models must be monitored after deployment.

20. الخلاصة

- أخلاقيات الذكاء الاصطناعي جزء من المسؤولية الهندسية.
- يمكن أن تنتج البيانات المنحازة نماذج منحازة.
- قابلية التفسير مهمة في التطبيقات عالية التأثير.
- لا يلغي الذكاء الاصطناعي المسؤولية البشرية.
- يجب مراعاة الخصوصية والأمن السيبراني.
- يجب أن يعمل AI ضمن قيود سلامة واضحة.
- وجود نظام احتياطي وإشراف بشري ضروري.
- يجب مراقبة النموذج بعد تشغيله.

📝 Review Questions

1. Why is AI ethics important in engineering?
2. What is data bias?
3. What is dataset shift?
4. What is explainable AI?
5. What is automation bias?
6. Why is human oversight important?
7. Give three cybersecurity risks.
8. Why is AI safety important in power systems?
9. What is a fail-safe mechanism?
10. Why must deployed models be monitored?

📝 أسئلة المراجعة

1. لماذا تعد أخلاقيات الذكاء الاصطناعي مهمة هندسيًا؟
2. ما المقصود بتحيز البيانات؟
3. ما المقصود بتغير توزيع البيانات؟
4. ما المقصود بالذكاء الاصطناعي القابل للتفسير؟
5. ما هو التحيز نحو الأتمتة؟
6. لماذا يعد الإشراف البشري مهمًا؟
7. اذكر ثلاثة مخاطر سيبرانية.
8. لماذا تكون سلامة AI مهمة في نظم الطاقة؟
9. ما المقصود بنظام Fail-Safe؟
10. لماذا يجب مراقبة النموذج بعد تشغيله؟

References | المراجع

المراجع العامة المستخدمة في هذا المقرر

[1] A. Gabash, *Flexible Optimal Operations of Energy Supply Networks with Renewable Energy Generation and Battery Storage*. Saarbrücken, Germany: Südwestdeutscher Verlag für Hochschulschriften, 2014.

[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th Global ed. Harlow, U.K.: Pearson Education Limited, 2022.

[3] A. Gabash, “Energy Market Transition and Climate Change: A Review of TSOs-DSOs C+++ Framework from 1800 to Present,” *Energies*, vol. 16, no. 17, Art. no. 6139, 2023, doi: 10.3390/en16176139.

[4] A. Gabash, *AI Applications in Electrical Power Systems*, Version 1.0, 2026. [Online]. Available: https://aoussgabash.com

Recommended citation:
Gabash, A. (2026). AI Ethics. AI Applications in Electrical Power Systems, Lecture 08, Version 1.0. https://aoussgabash.com

المرجع المقترح:
أوس غباش، 2026، محاضرة 08، تطبيقات الذكاء الاصطناعي في أنظمة الطاقة الكهربائية، الإصدار 1.0، https://aoussgabash.com

© 2026 Dr.-Ing. Aouss Gabash. Educational use with attribution to the author and official website.