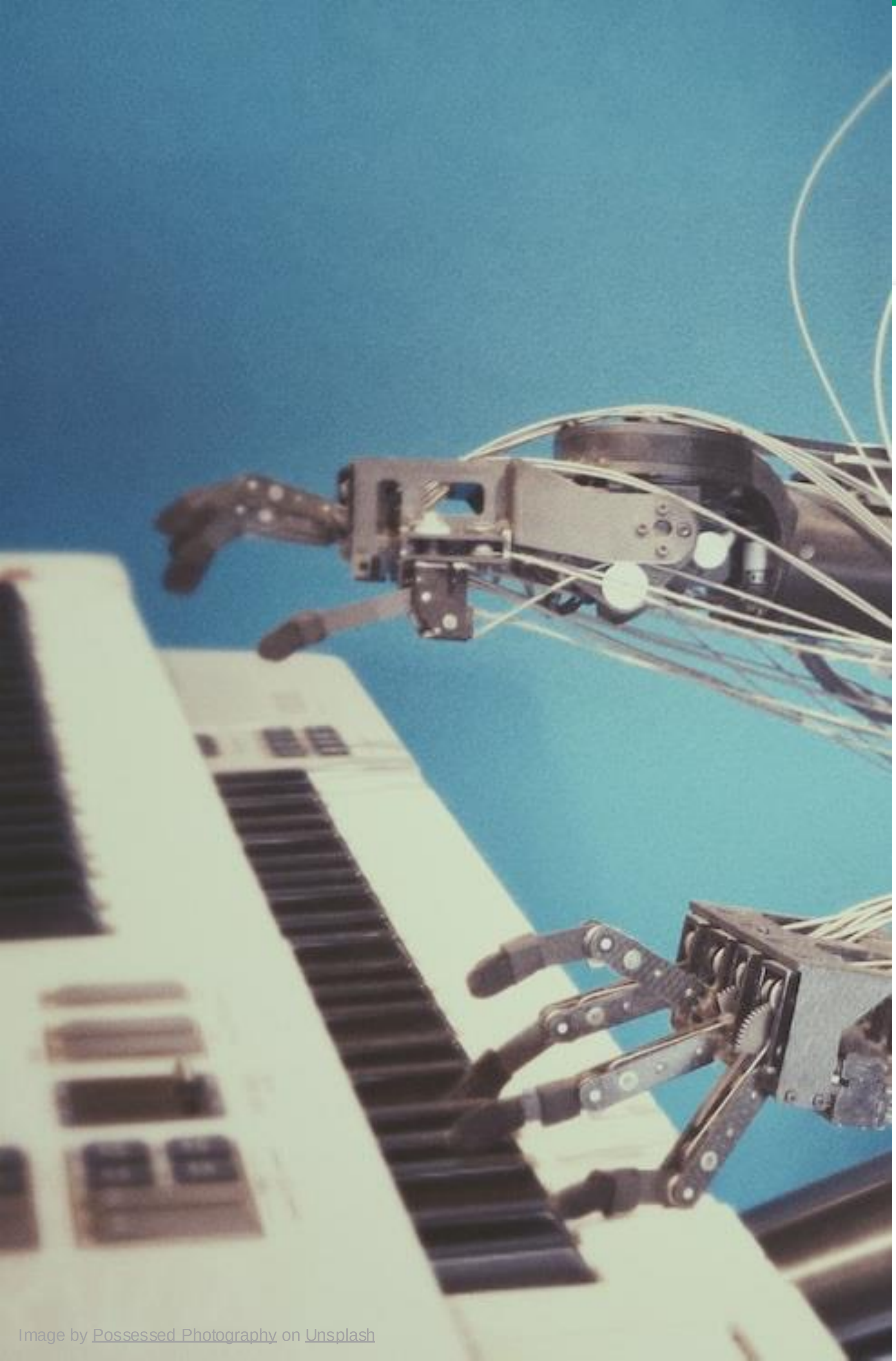


Image by Possessed Photography on Unsplash

Applied Artificial Intelligence
08 - Human-AI Collaboration

Univ.-Prof. Dr-Ing. habil. Niklas Kühl
www.niklas.xyz

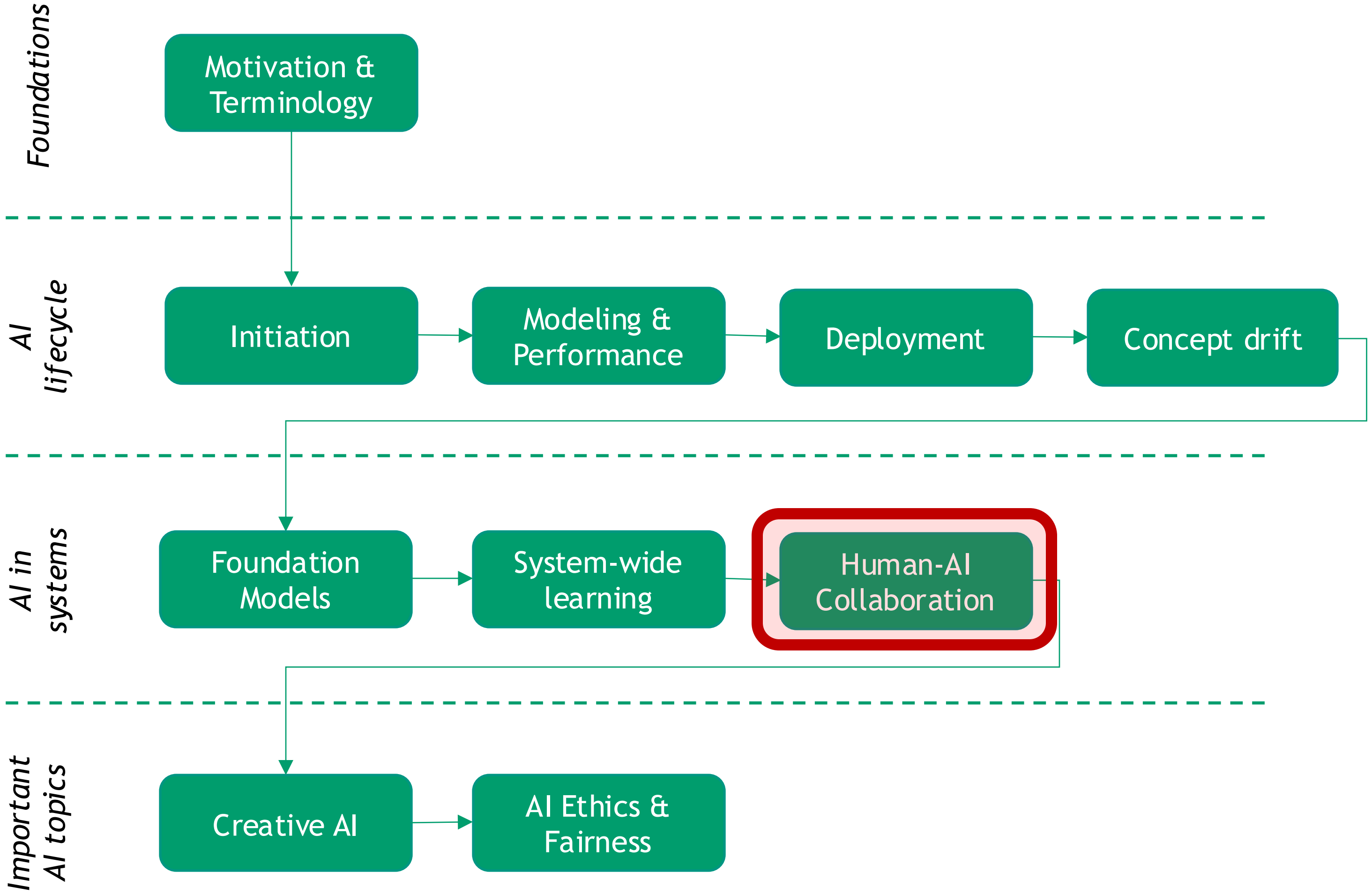
University of Bayreuth

Karlsruhe Institute of Technology

TUM School of Management

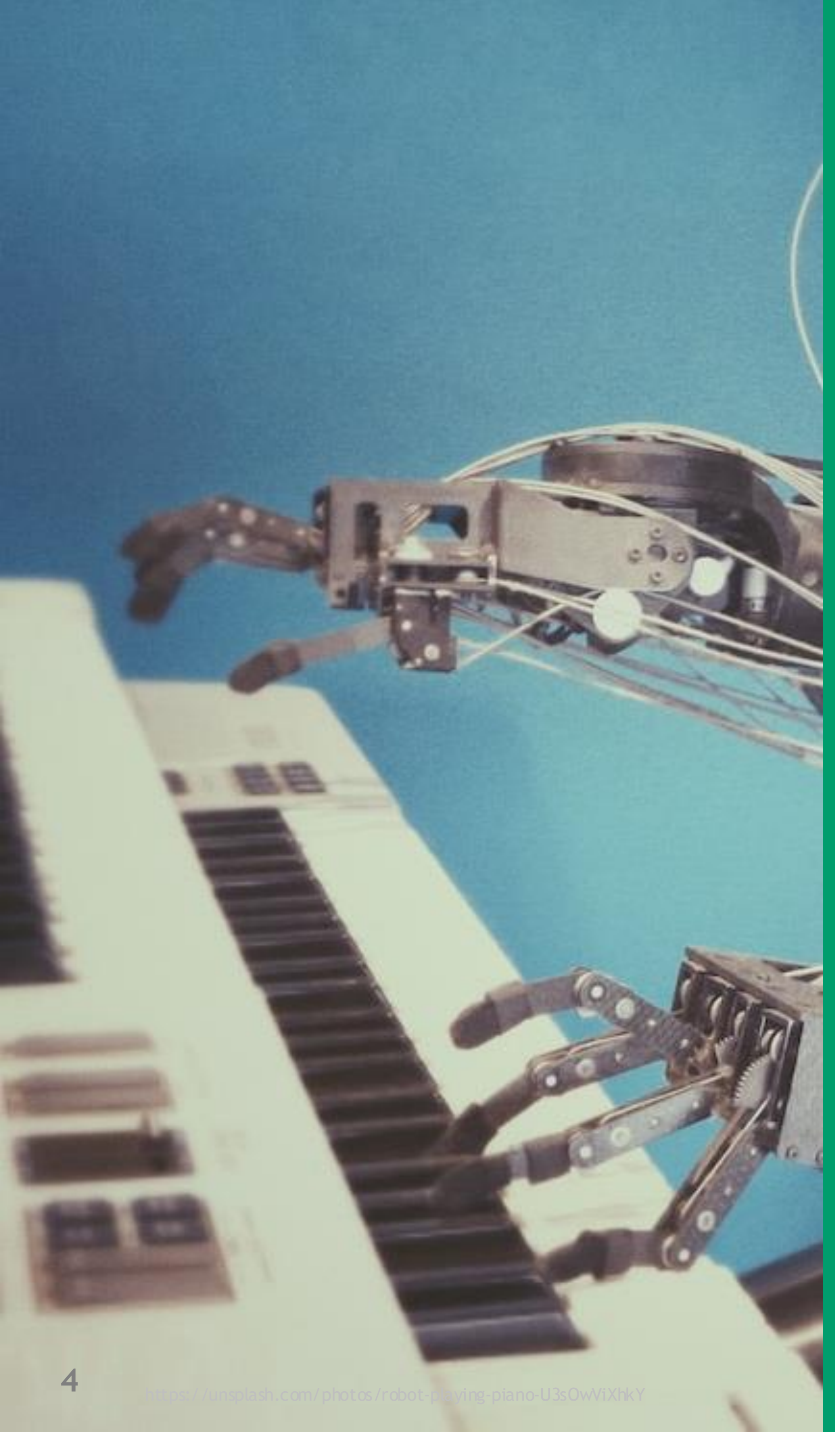
www.uni-bayreuth.de | www.kit.edu | www.tum.de | www.fim-rc.de | www.wirtschaftsinformatik.fraunhofer.de

Organizational
The story of the lecture



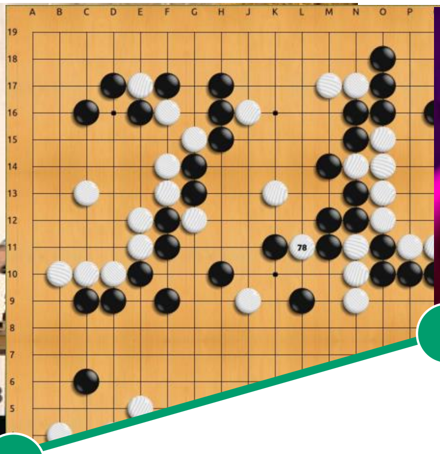
Objectives
What are the learning goals of this lecture?

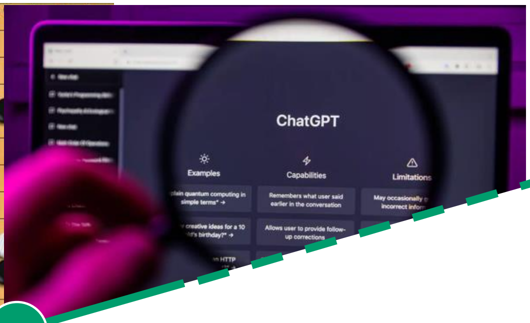
EXPLORE	UNDERSTAND	INTENSIFY	APPLY
Explore what the idea of Human-AI collaboration covers and why it is practically relevant	Understand how humans can complement and rely on AI	Get to know different mechanisms of Human-AI Collaboration	Apply the concepts of uncertainty quantification and explainability to AI artifacts



https://unsplash.com/photos/robot-playing-piano-U3sOwViXhkY

Introduction to Human-AI Collaboration
Where we come from and where we (might) go





1769
Workers vs. Steam Engine

1997
Kasparov vs. Deep Blue

2017
Lee Sedol vs. AlphaGo & Zero

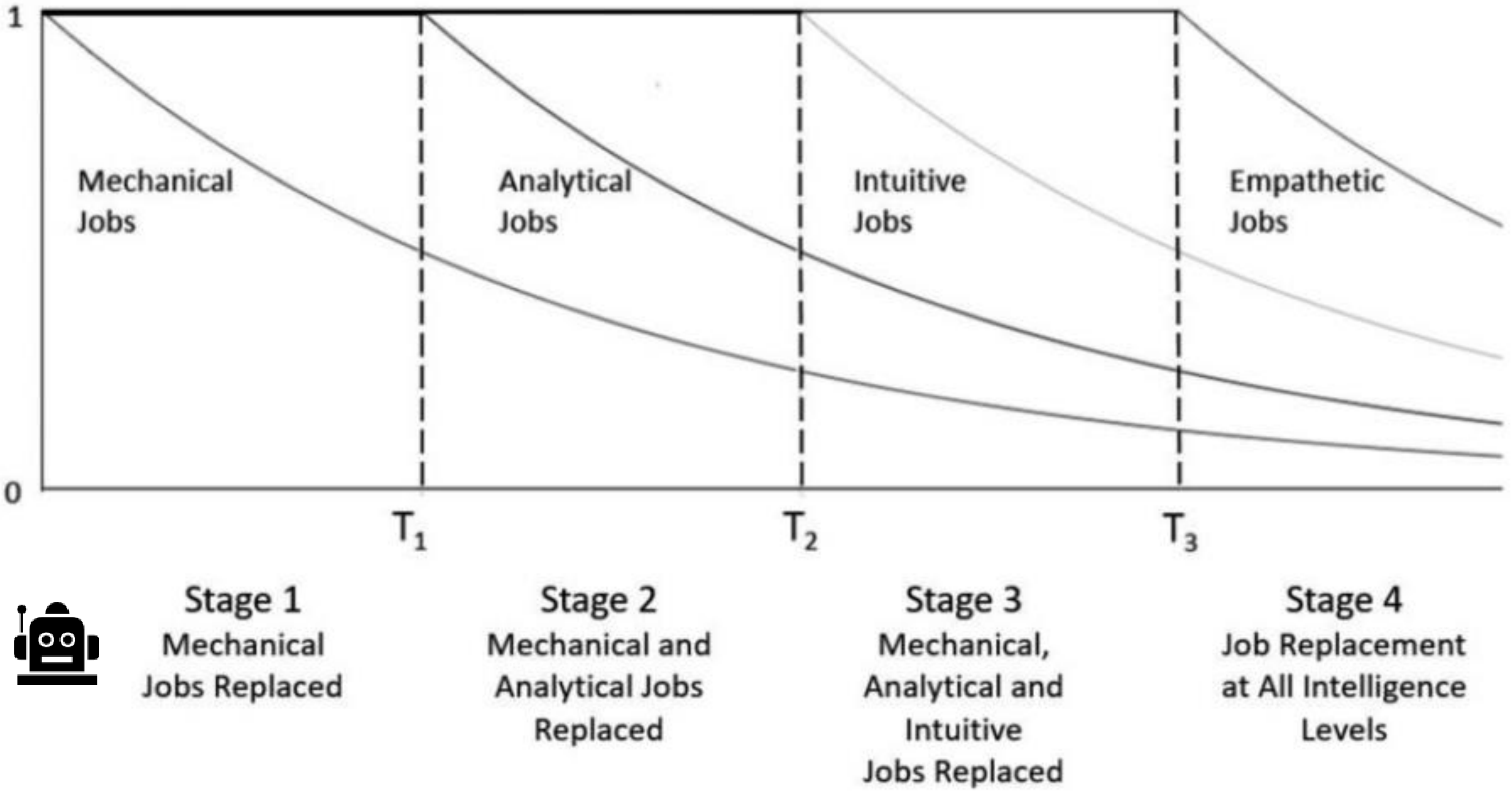
202X
Today

There is a 50% chance that "unaided machines can accomplish every task better and more cheaply than human workers" within the next 45 years [1].

By 2055 "half of today's work activities could be automated" [2].

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729-754. [1]
Manyika, J., Chui M., Miremadi, M., Bughin, J., George, K., Willmott, P., Dewhurst, M. (2017) A Future That Works: Automation, Employment and Productivity. McKinsey & Company, New York. [2]
Images: https://live.staticflickr.com/7009/6420453543_015e316461_b.jpg; https://upload.wikimedia.org/wikipedia/commons/d/d5/Kasparov-11.jpg, https://upload.wikimedia.org/wikipedia/commons/0/03/Lee-sedol-alphago-divine-move.jpg, https://upload.wikimedia.org/wikipedia/commons/6/61/Image-chatgpt.webp

Introduction to Human-AI Collaboration
Will AI take our jobs?



Huang, M., Rust, R. (2018). Artificial Intelligence in Service. Journal of Service Research, 21(1).

Huang and Rust (2018)

Introduction to Human-AI Collaboration
What do we want AI to do, and what do we want keep doing ourselves?

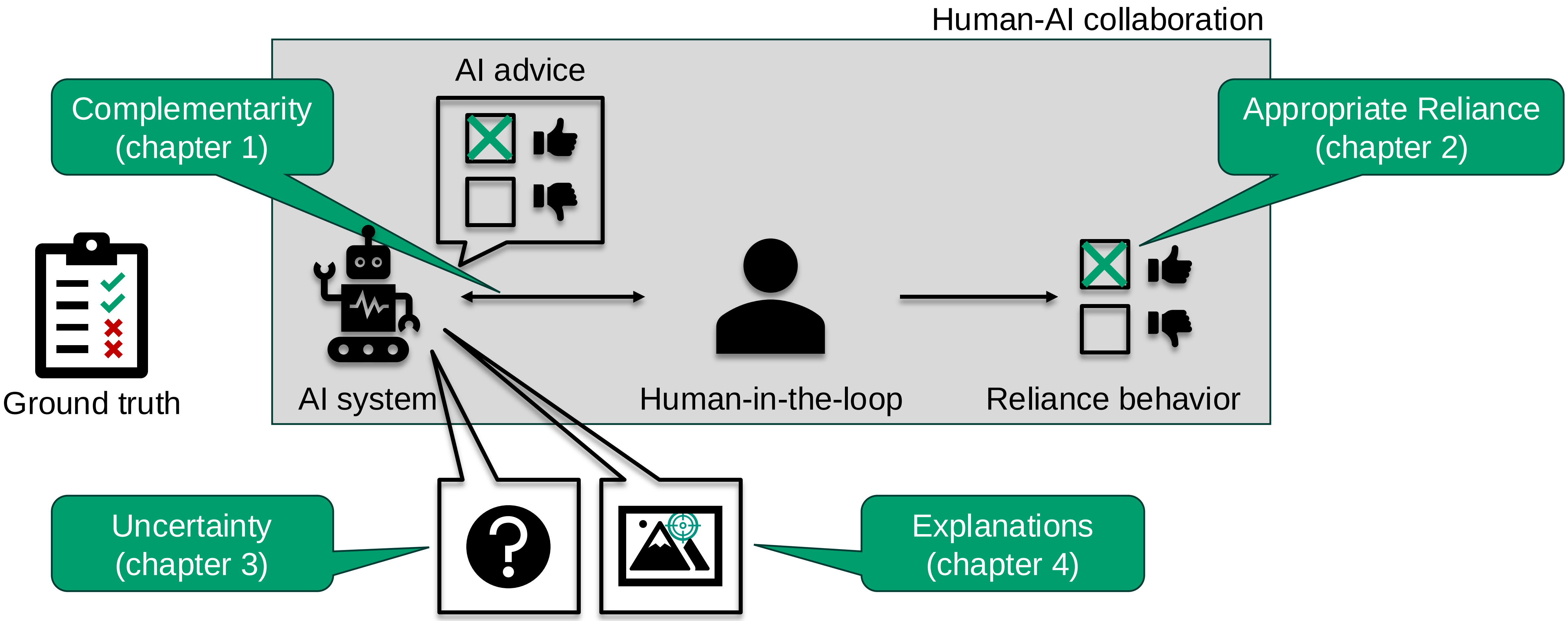
Difficulty

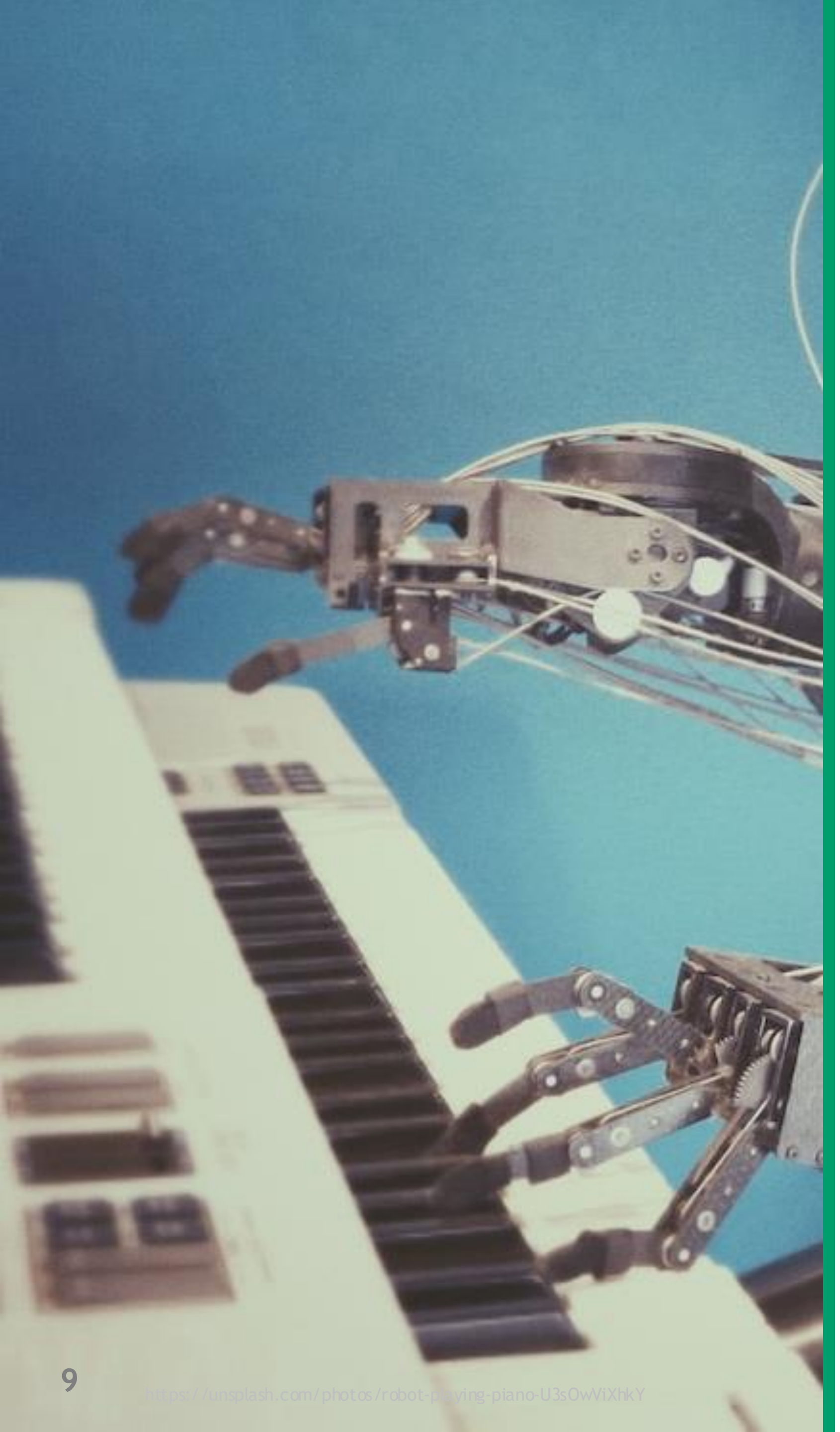
1. Automating what humans don’t want to do.	E.g., documentation, scribing, protocol compliance
2. Scaling what humans practically can’t do.	E.g., richer documentation, continuous assessments, long-term monitoring, continuous cognitive assistance, managing complexity
3. Performing what humans used to do at a “superhuman” level.	(In the Future? …LLMs? AGI?)

Yeung K. (2020). Recommendation of the Council on Artificial Intelligence (OECD). International Legal Materials, 59(1), 27-34. [1]

Introduction to Human-AI Collaboration

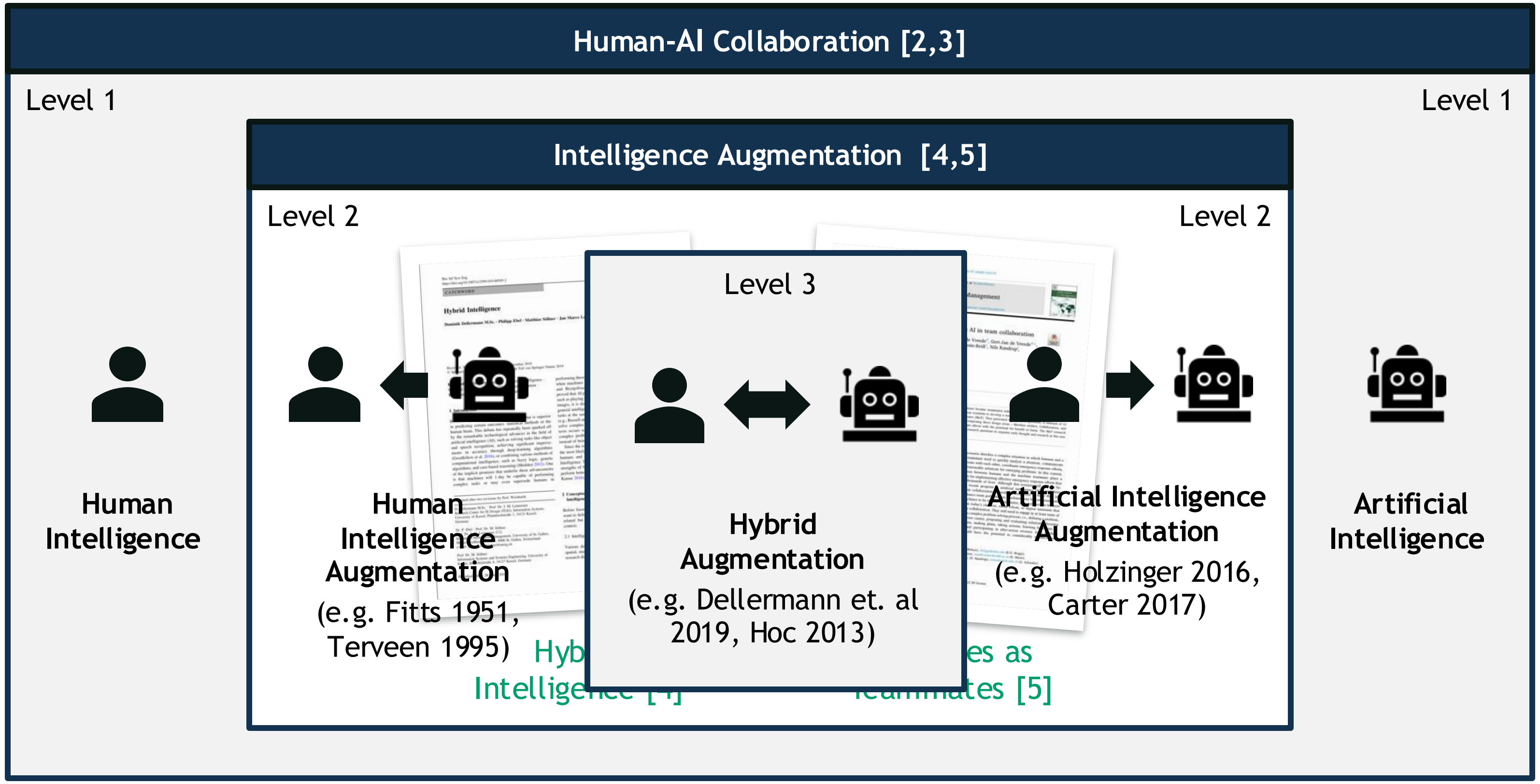
We will require some key terms and concepts today





https://unsplash.com/photos/robot-playing-piano-U3sOwViXhkY

Modes of Collaboration
Generally, we aim to collaborate by leveraging complementary capabilities…



Krämer, N., Simons, N., Kopp, S. (2007). The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry. Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, 238251. [1]
Silverman (1992). Evaluating and Refining Expert Critiquing Systems: A Methodology. Decision Science, 23(1), 86-110. [2]
Terveen, L.G. (1995). Overview of human-computer collaboration. Knowl. Based Syst., 8, 67-81. [3]
Dellermann, D., Ebel, P., Leimeister, M., Söllner, M. (2019). Hybrid Intelligence. Bus Inf Syst Eng 61, 637-643; [5] Seeber, I., Bittner, E., Briggs, R. (2019). Machines as teammates: A research agenda on AI in team collaboration. Information & Management, 1-22. [4]
Voessing (2020). Designing human-computer collaboration: Transparency and automation for intelligence augmentation. KIT. [5]
Image: XKCD

Complementarity
…but, as with humans, collaboration is not always so easy

Image created with Midjourney

Complementarity
…but, as with humans, collaboration is not always so easy

Forbes

Lawyer Used ChatGPT In Court —And Cited Fake Cases. A Judge Is Considering Sanctions [1]

The Guardian

Canada lawyer under fire for submitting fake cases created by AI chatbot [2]

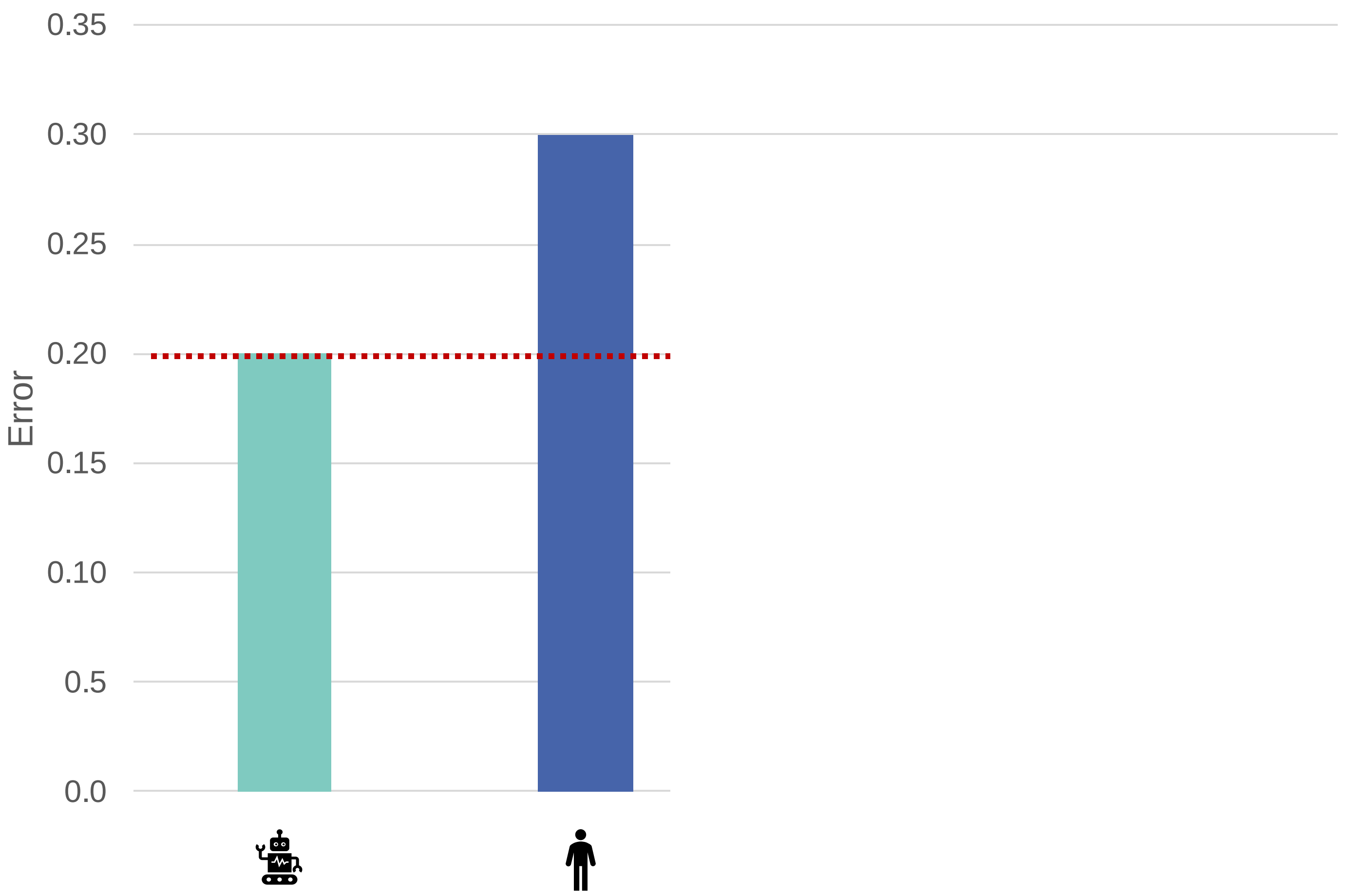
[3]

https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/ [1]
https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/ [2]
Created with Midjourney [3]

C

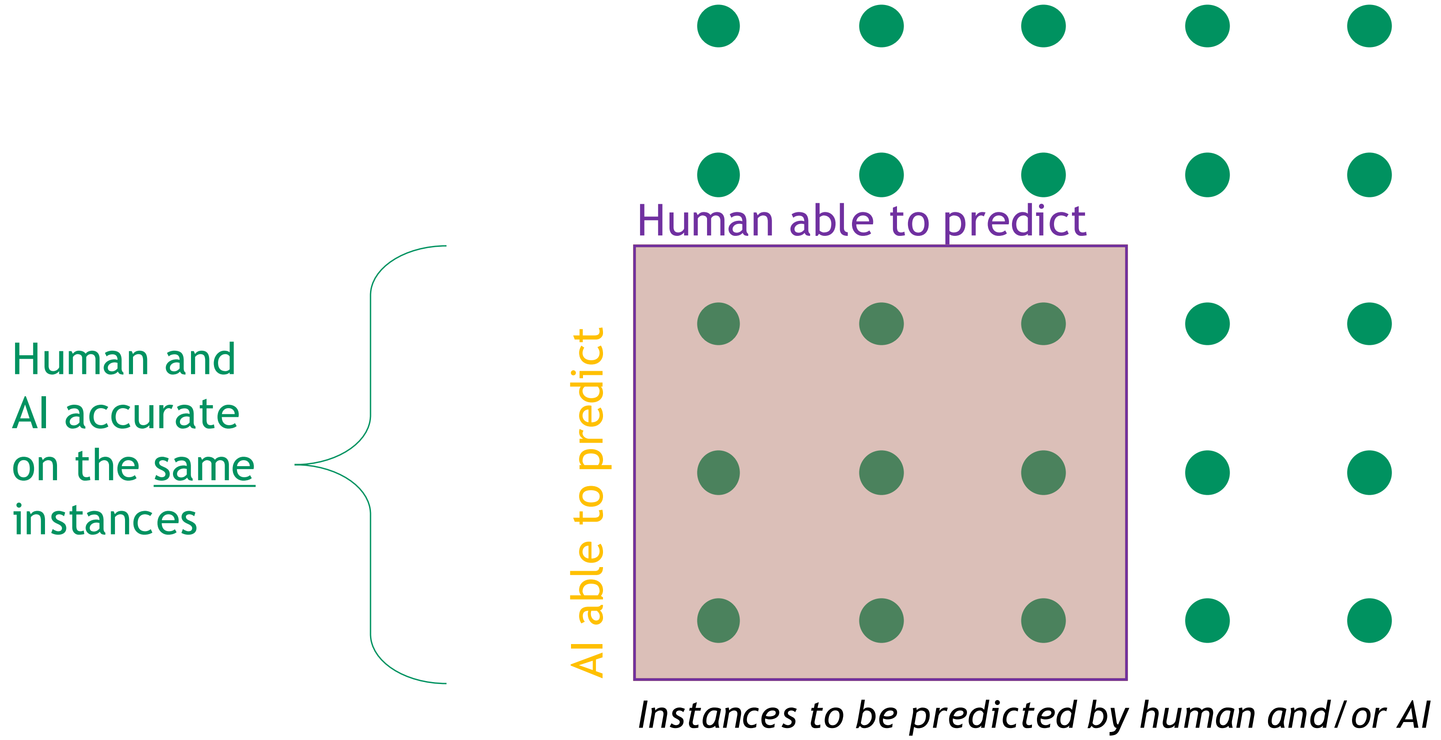
Complementary Team Performance

The potential depends on the performance of both entities…



Performance (CTP)

Complementary Potential
…as well as the distribution of the individual strengths (1/3)



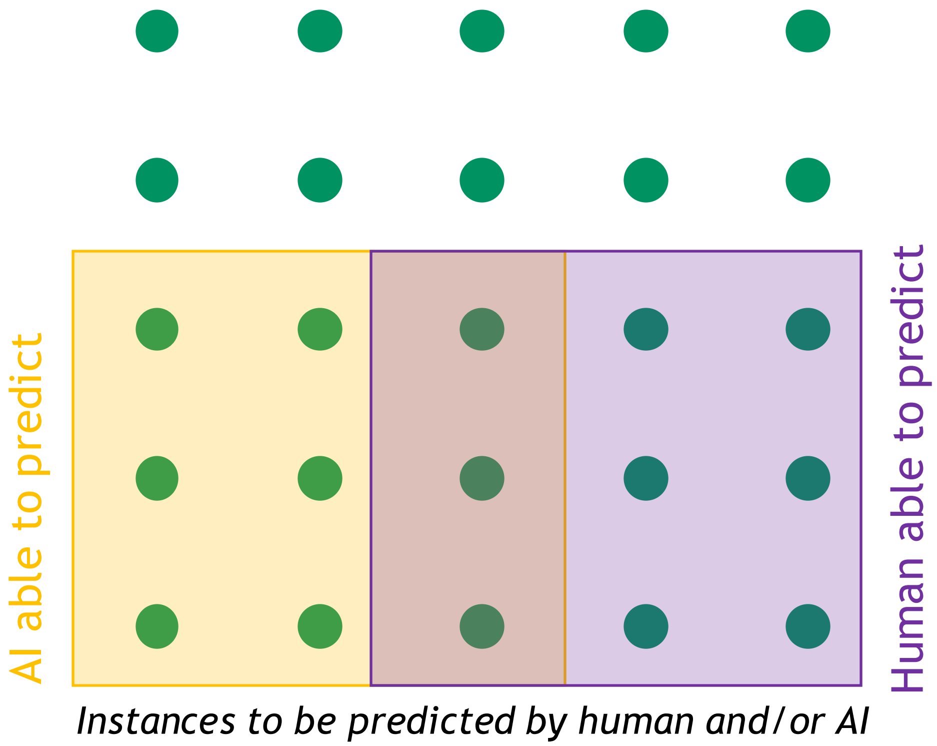
= No Complementary Potential

→ No Complementary Team Performance possible

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., Satzger, G. (2022). On the Effect of Information Asymmetry in Human-AI Teams. CHI Conference on Human Factors in Computing Systems (CHI '22), ACM CHI Workshop on Human-Centered Explainable AI

Complementary Potential
…as well as the distribution of the individual strengths (2/3)

Human and AI accurate on <u>different</u> instances



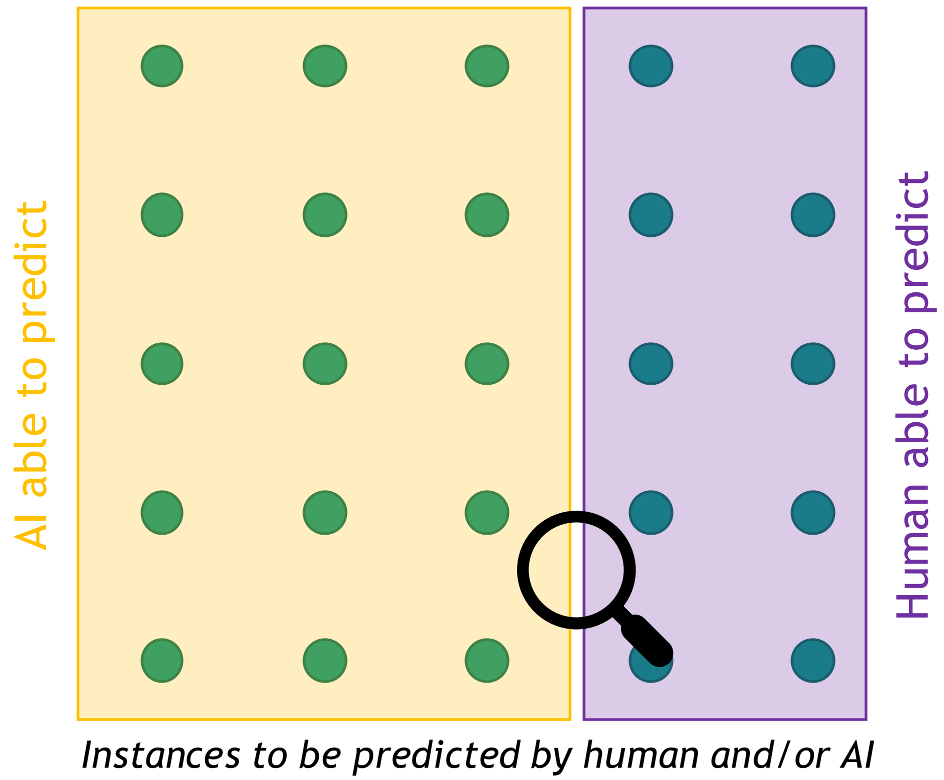
= Complementary Potential

→ Complementary Team Performance possible

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., Satzger, G. (2022). On the Effect of Information Asymmetry in Human-AI Teams. CHI Conference on Human Factors in Computing Systems (CHI '22), ACM CHI Workshop on Human-Centered Explainable AI

Complementary Potential
…as well as the distribution of the individual strengths (3/3)

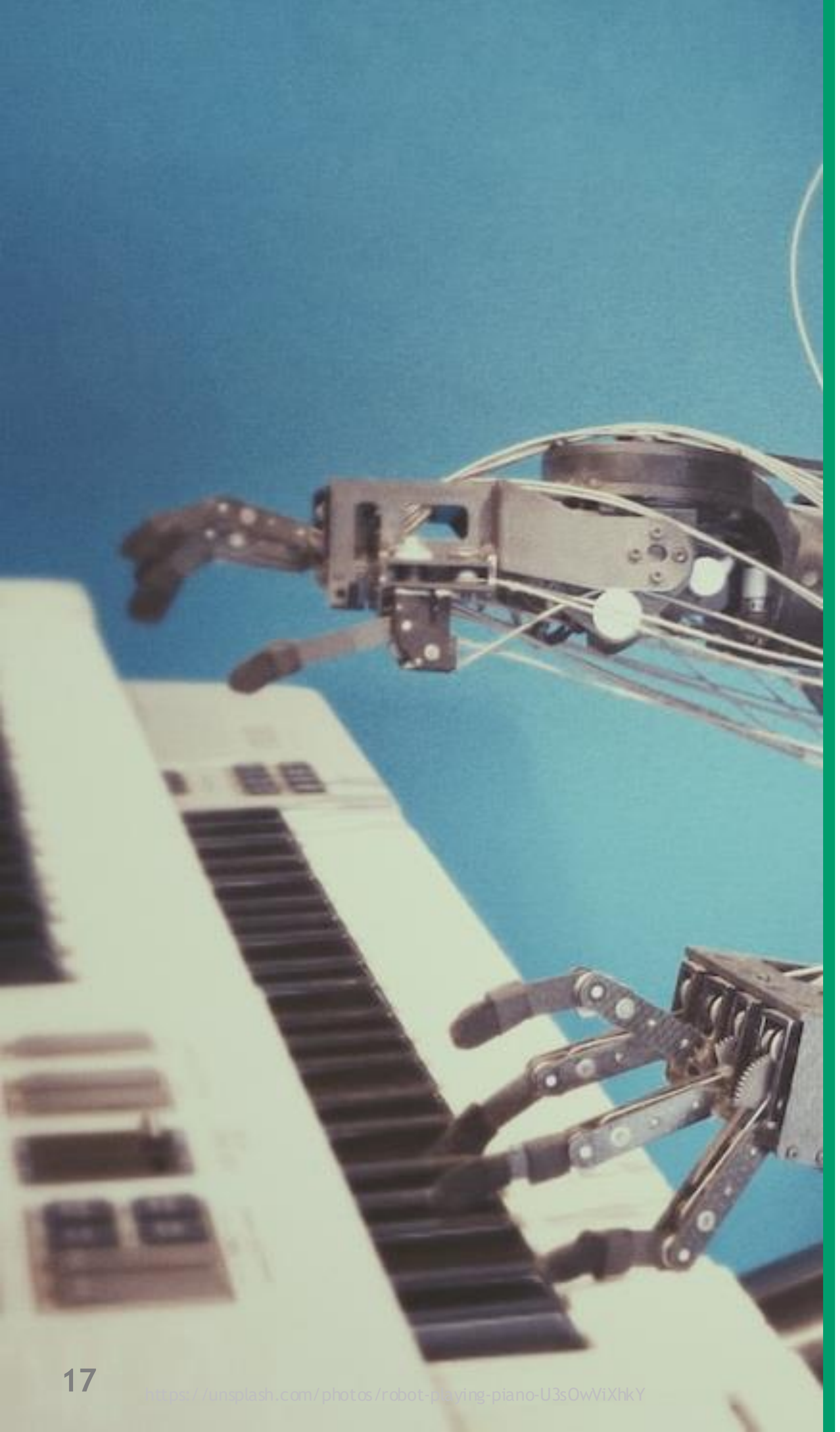
Human and AI accurate on <u>entirely different</u> instances



= **<u>Maximum</u> Complementary Potential**

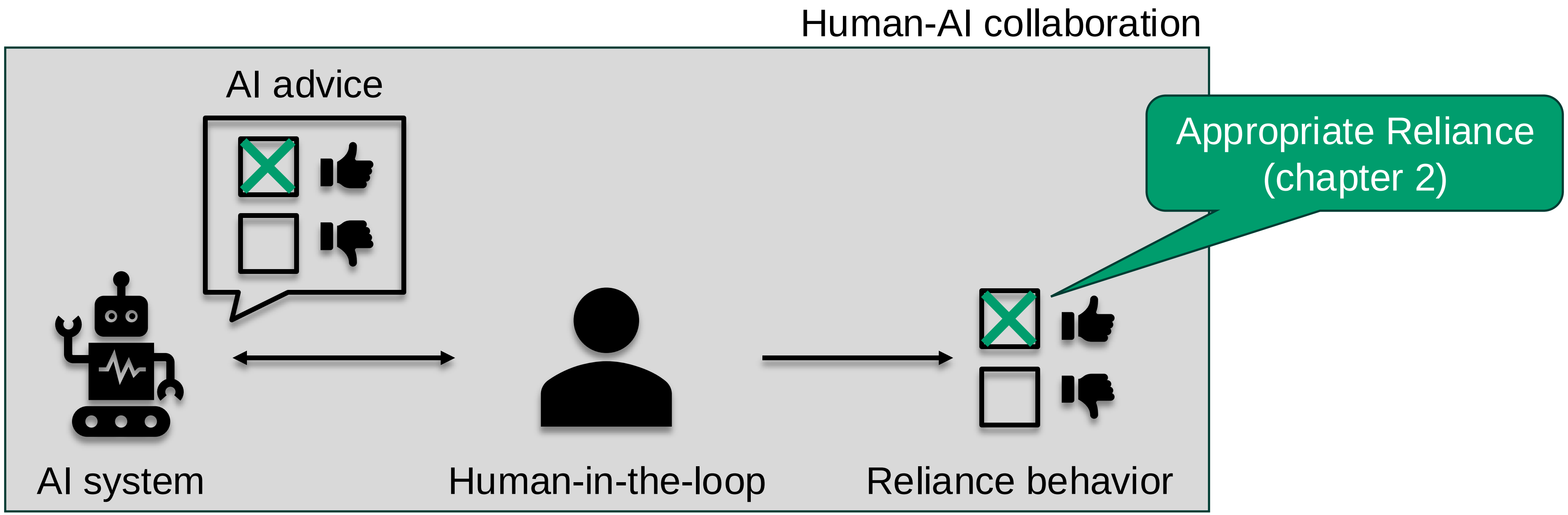
→ Complementary Team Performance possible

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., Satzger, G. (2022). On the Effect of Information Asymmetry in Human-AI Teams. CHI Conference on Human Factors in Computing Systems (CHI '22), ACM CHI Workshop on Human-Centered Explainable AI



https://unsplash.com/photos/robot-playing-piano-U3sOwViXhkY

Appropriate Reliance



Overreliance: Automation Bias
Humans can experience “automation bias” when using AI

- According to Monsier and Skitka, **automation bias** is "the tendency for individuals to over-rely on automated systems, to the exclusion of other information or their own decision-making skills”.
- Humans often have a tendency to trust the decisions and actions of AI systems **(a) without fully considering other options** or **(b) using their own judgement**.
- This can lead to problems such as over-reliance on the AI system and a lack of critical thinking on the part of the human.

Reasons for automation bias:

- Perceived reliability and accuracy of the AI system
- Assumption that the AI system is objective and unbiased
- Lack of transparency and understanding of the AI system's decision-making processes
- Ease and convenience of relying on the AI system rather than using one's own judgment.

Image: https://www.flickr.com/photos/xingxiyang/13568821165/

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), Automation and human performance: Theory and applications (pp. 201-220). Lawrence Erlbaum Associates.

Underreliance: Algorithm Aversion

Contrarily, humans can experience “algorithm aversion”

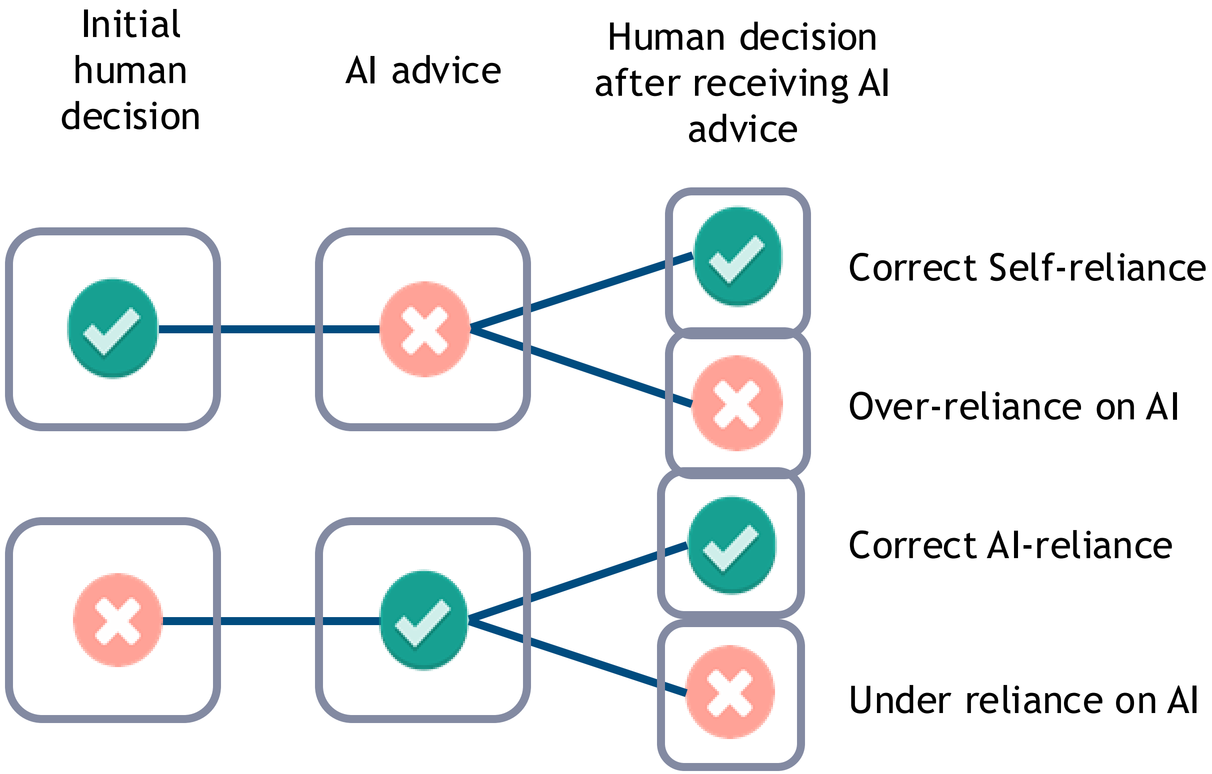
- According to Jussupow et al., **algorithm aversion** is a "biased assessment of an algorithm which manifests in negative behaviours and attitudes towards the algorithm compared to a human agent”.
- Humans often have a tendency to **assess algorithmic output less favorably than human output**, even if they are identical
- This can lead to problems such as over-reliance on flawed human decisions just because they are human.

Reasons for algorithmic averison:

- Perceived performance and capabilities of the AI system
- Lack of human agency and involvement
- Preference of human expertise (e.g., experienced physicians)
- Preference of socially closer human agents (e.g., friends)
- Distrust in technology

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.
Images: Freepik

Conceptualizing collaborative decisions
Appropriate reliance means relying when right and overruling when wrong



Schemmer, M., Kuehl, N., Benz, C., Bartos, A., Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23). Association for Computing Machinery, New York, NY, USA, 410-422.

Conceptualizing collaborative decisions

Appropriate reliance is the “sweet spot” between the “extremes”

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. [1]
Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association. [2]
Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023, March). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces. [3]

Conceptualizing collaborative decisions
Appropriate reliance is crucial when human oversight is (legally) demanded


Human oversight not good enough for AI war machines

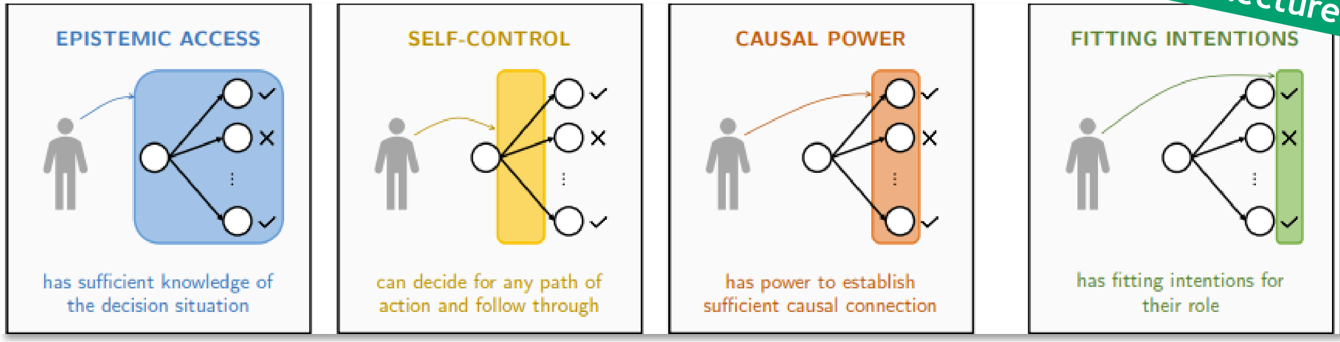
Humans put excessive trust in machine systems and their conclusions, raising risk they view real people as mere items on a screen

By MARK TSAGAS
AUGUST 29, 2024

[1]

- Many high-stakes domains can or should not be fully automated:
  - Recruiting
  - Legal judgments
  - Warfare
  - …
- In these cases, **human oversight** is legally or technically required
- Effective human oversight requires four conditions:

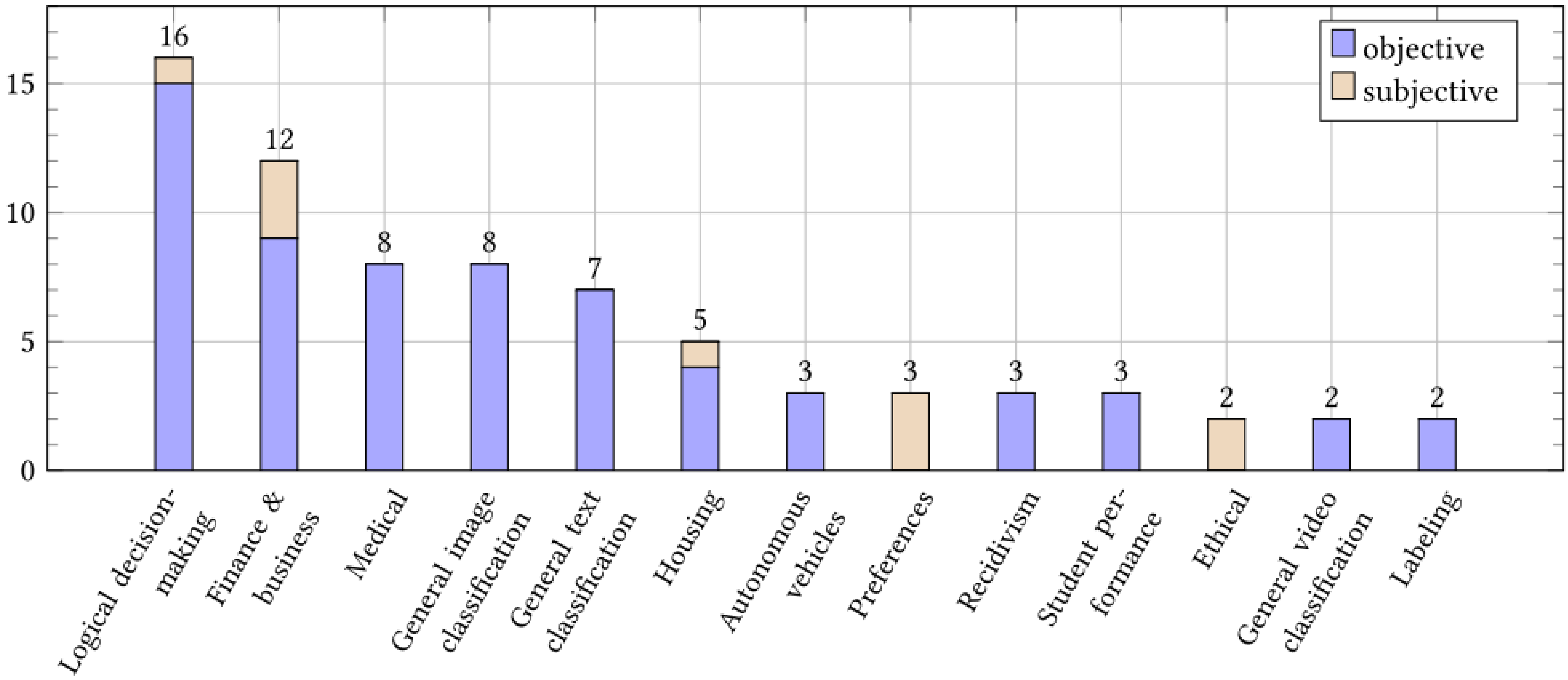
see lecture 10



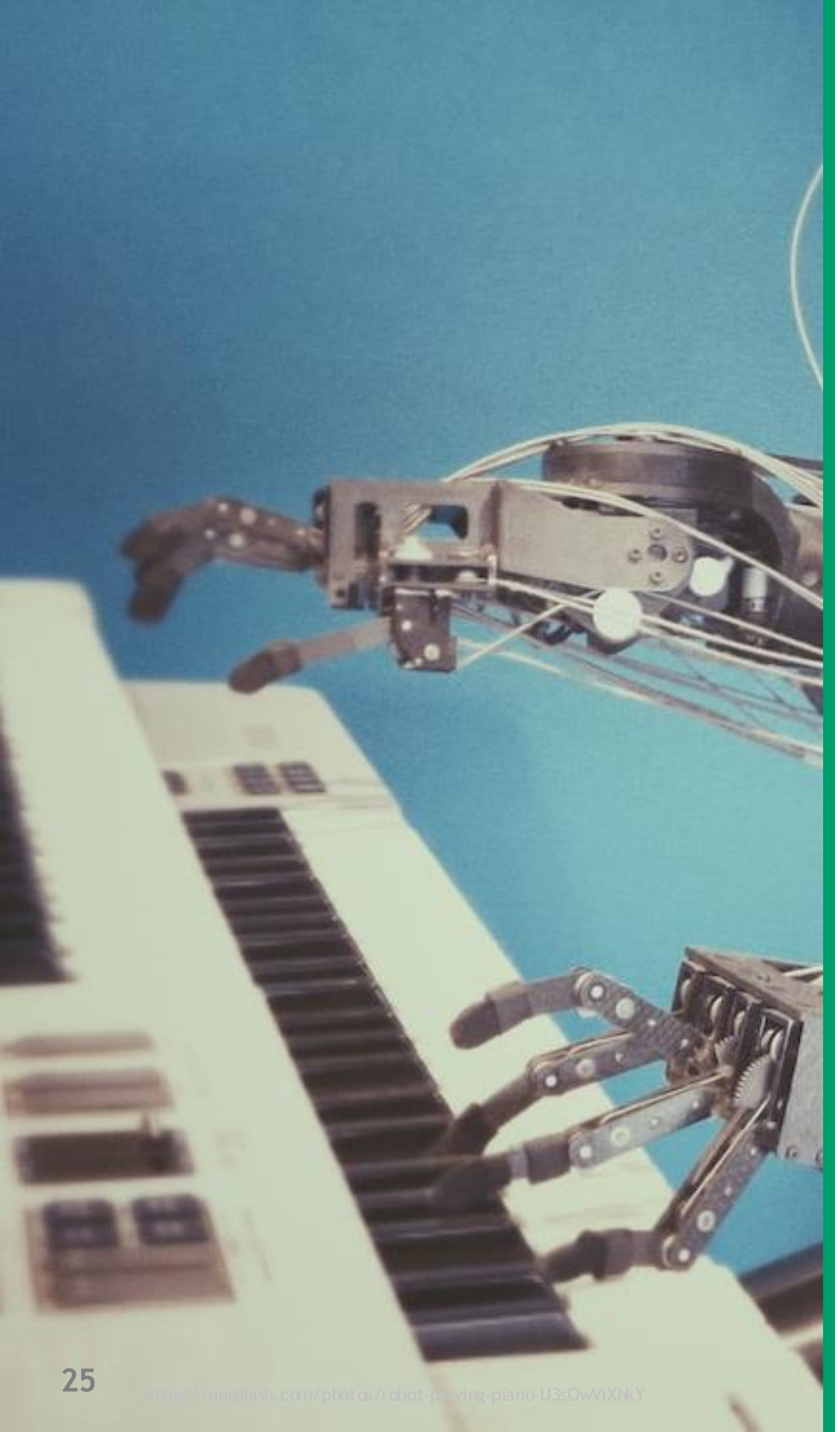
[2]

Asia Times, 29.08.2024, https://asiatimes.com/2024/08/human-oversight-not-good-enough-for-ai-war-machines/ [1]
Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024, June). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In The 2024 ACM Conference on Fairness, Accountability, and Transparency. [2]

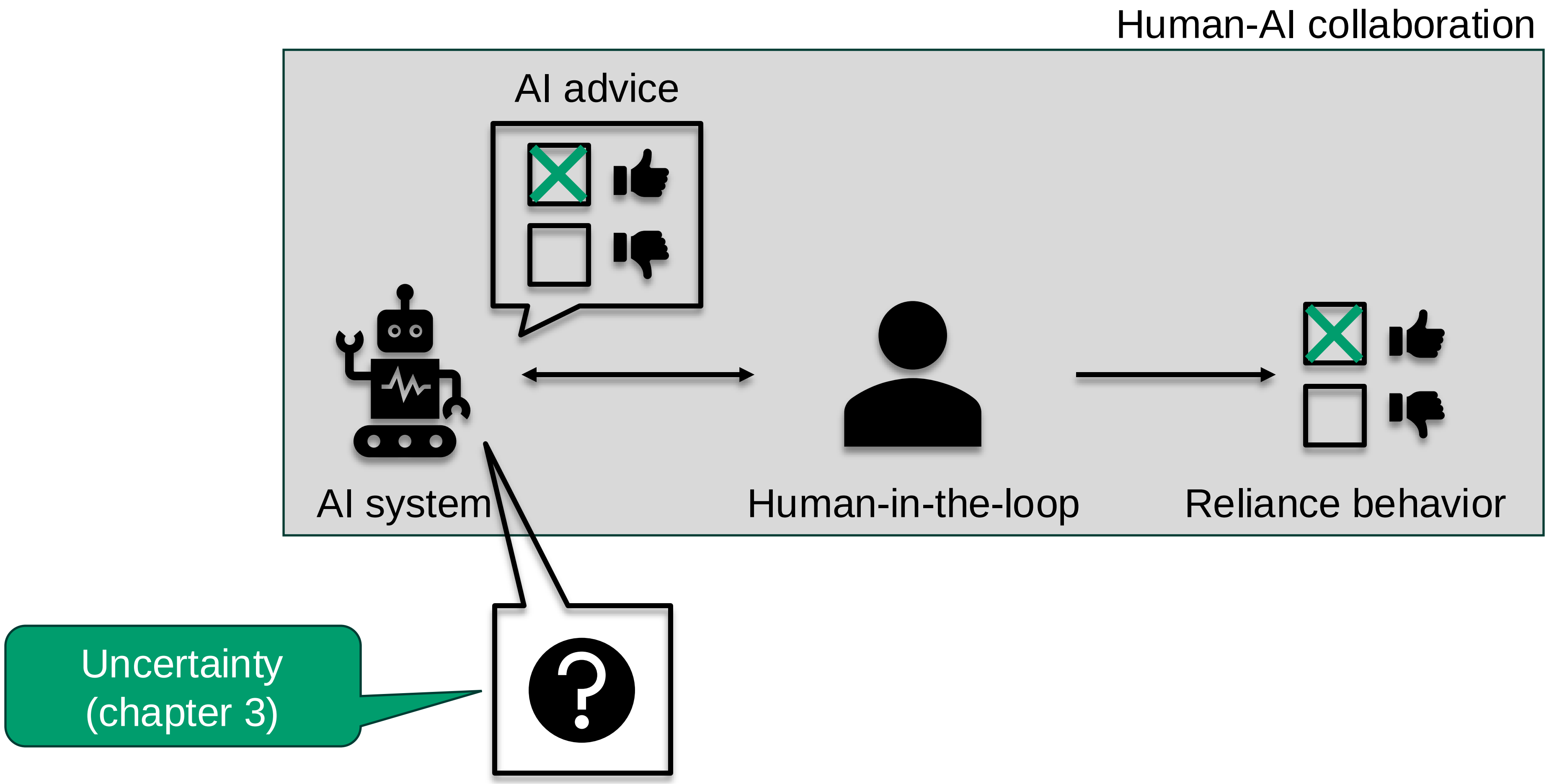
Appropriate Reliance | Research
Appropriate reliance is an ongoing field of research with broad applicability



Eckhardt, S.; Kühl, N.; Dolata, M.; Schwabe, G. (2024): A Survey of AI Reliance. Working paper.


https://unsplash.com/photos/robot-playing-piano-U3sOwViXhkY

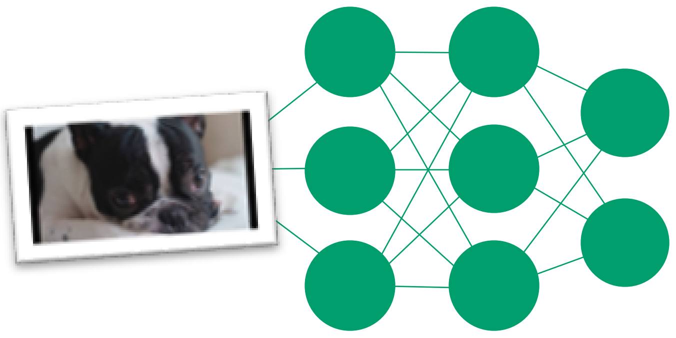
Uncertainty



Uncertainty
Uncertainty quantification might steer appropriate reliance

- **Epistemic uncertainty [1]:** “model uncertainty” due to limited data and knowledge.
- **Aleatoric uncertainty [1]:** “data uncertainty” due to inherent noise or “natural randomness” in the data.

“Gal and Ghahramani [2] have introduced a [...] simple method for capturing [...] uncertainty. They have discovered that training any NNs with **dropouts** [...] could be interpreted as an approximate inference of the weight’s posterior [...]. One simply needs to make multiple predictions with the trained model and average them.”

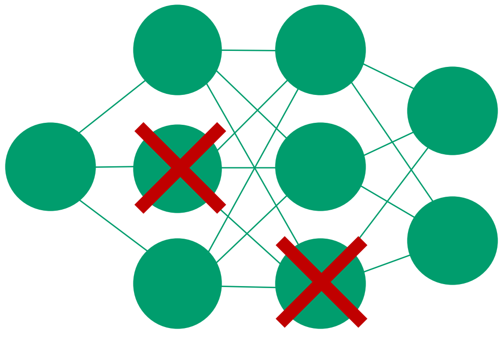


60% Dog

40% Cat

?

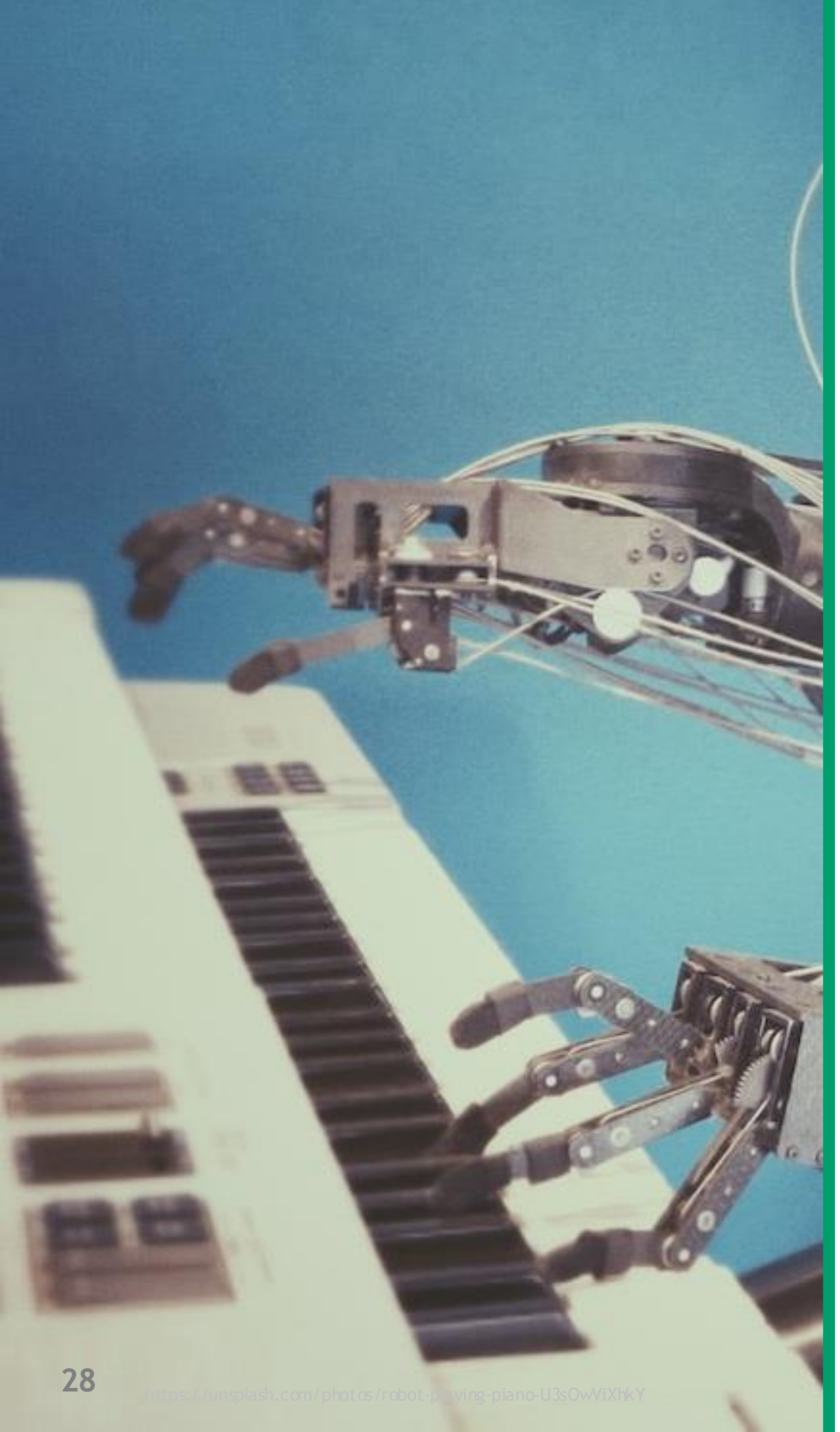
= Uncertainty



Ø = 45% Dog

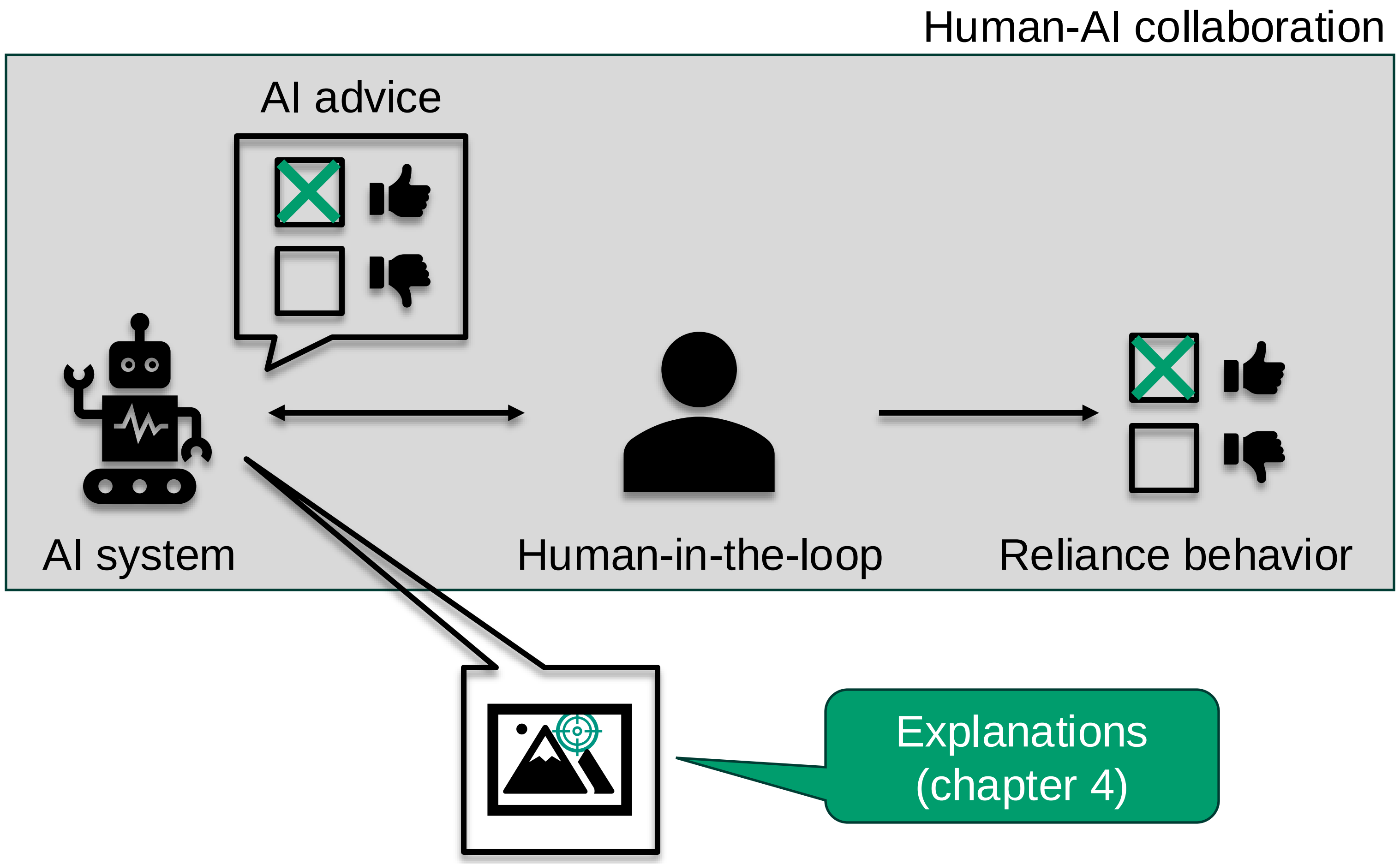
Ø = 55% Cat

Der Kiureghian, A., & Ditlevsen, O. D. (2009). Aleatoric or epistemic? Does it matter? Structural Safety, 31(2), 105-112. [1]
Gal, Y. & Ghahramani, Z.. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceedings of The 33rd International Conference on Machine Learning, in Proceedings of Machine Learning Research, 48, 1050-1059 [2]
Inovex (2020); DeepLearning.ai (2020)



https://unsplash.com/photos/robot-playing-piano-U3sOwViXhkY

Explanations

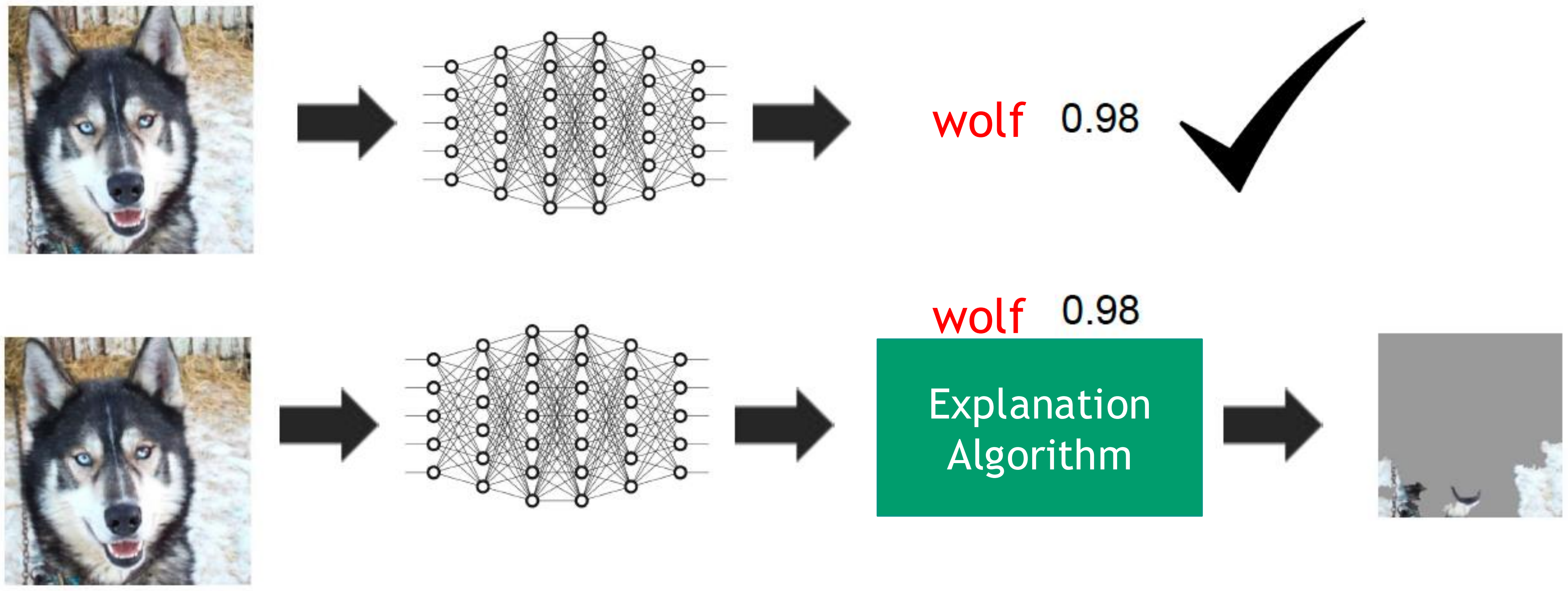


Explanations | Example
Wolf or husky? (1/3)

95% Accuracy.
Should we trust the model?

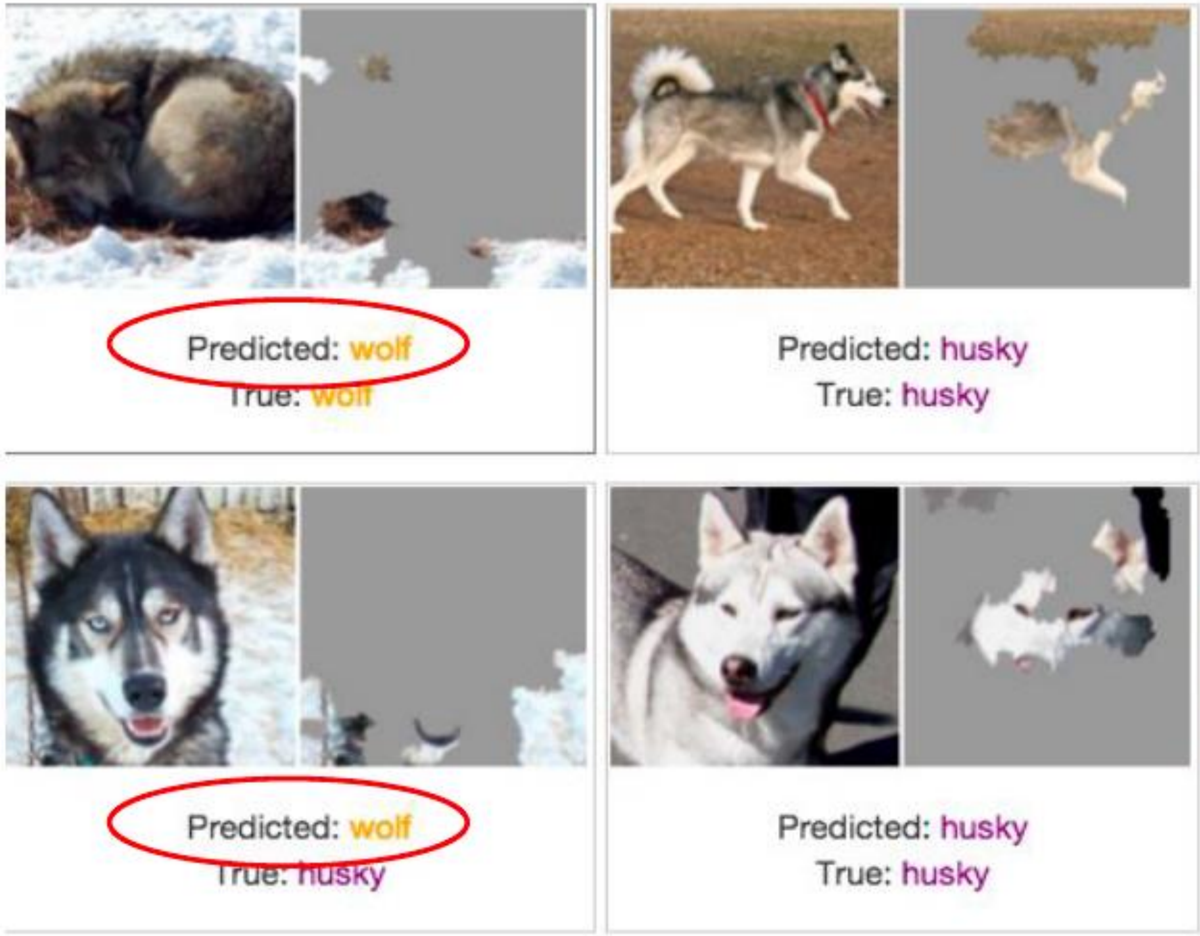
explainml-tutorial.github.io; Ribeiro, M., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135-1144.

Explanations | Example
Wolf or husky? (2/3)



explainml-tutorial.github.io; Ribeiro, M., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135-1144.

Explanations |Example
Wolf or husky? (3/3)



We built a great snow detector…

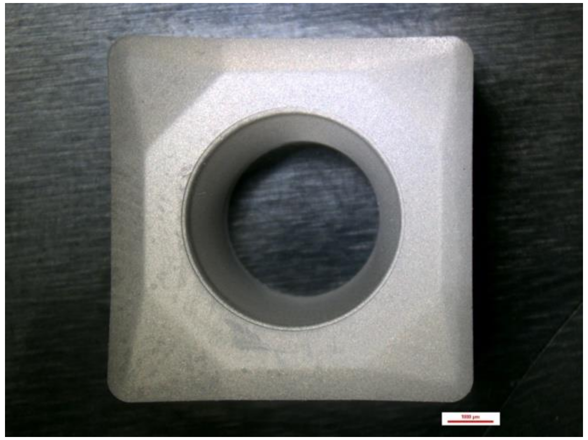
explainml-tutorial.github.io; Ribeiro, M., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135-1144.

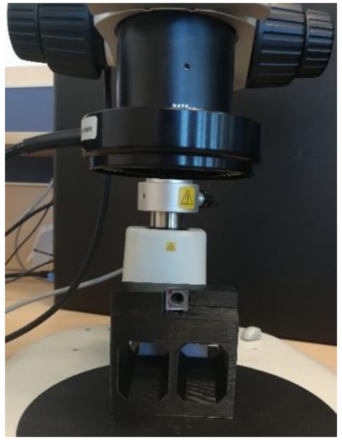
Explanations | Example (1/3)

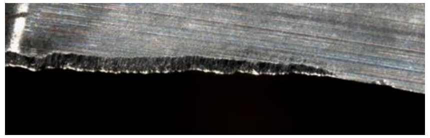
Making wear analysis in the manufacturing industry more inefficient

Our Research

- **Wear analysis is essential for…**
  - improving machining processes of customers
  - developing new generations of cutting tools
- **Objectives:**
  - Automatically characterize wear on machining tools
  - Provide supplementary “data-based service” to customers

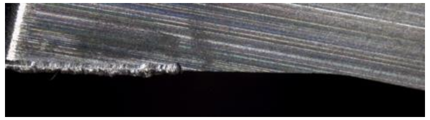






Flank wear

Chipping



Built-up edge

No wear

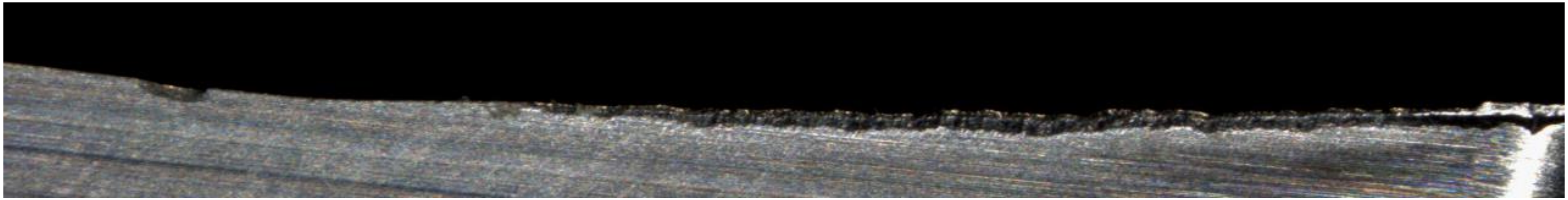
Walk, J., Kühl, N., Schäfer, J. (2020). Towards Leveraging End-of-Life Tools as an Asset: Value Co-Creation based on Deep Learning in the Machining Industry. Hawaii International Conference on System Sciences;
Treiss, A., Walk, J., Kühl, N. (2021). An Uncertainty-Based Human-in-the-Loop System for Industrial Tool Wear Analysis. In: Dong, Y., Ifrim, G., Mladenić, D., Saunders, C., Van Hoecke, S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science, Springer.

Explanations | Example (2/3)

First, we need a pixel-wise classification of different types of wear

Our Research

152 images…



…with pixel-wise annotation.

Chipping   Flank Wear   Build-Up Edge

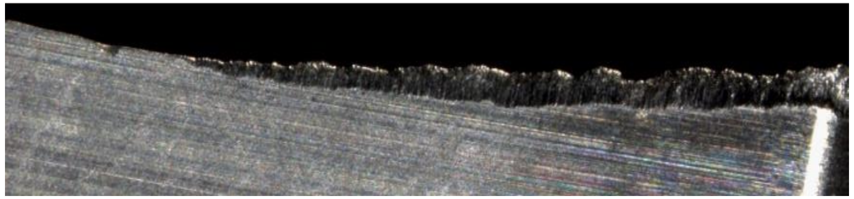
Walk, J., Kühl, N., Schäfer, J. (2020). Towards Leveraging End-of-Life Tools as an Asset: Value Co-Creation based on Deep Learning in the Machining Industry. Hawaii International Conference on System Sciences;
Treiss, A., Walk, J., Kühl, N. (2021). An Uncertainty-Based Human-in-the-Loop System for Industrial Tool Wear Analysis. In: Dong, Y., Ifrim, G., Mladenić, D., Saunders, C., Van Hoecke, S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science, Springer.

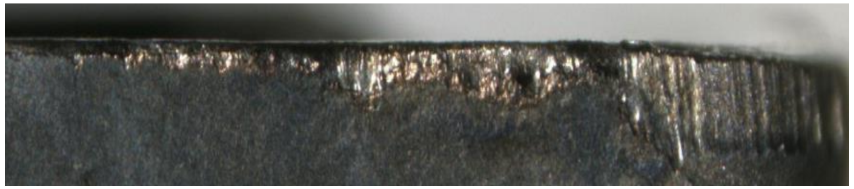
Explanations | Example (3/3)

Then, we add uncertainty quantification to direct human efforts

Our Research

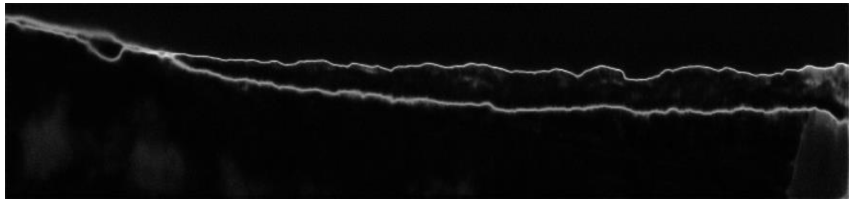
Input

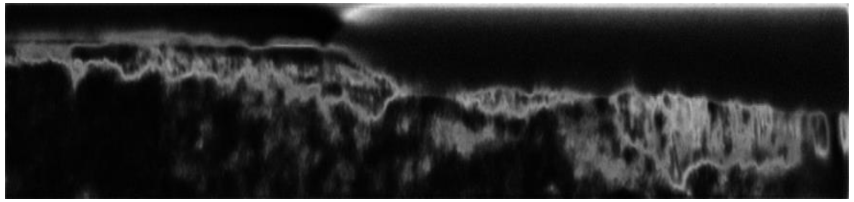




Output

Quantified Uncertainty





Valid output

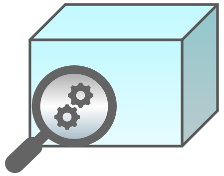
Manual labelling necessary

Walk, J., Kühl, N., Schäfer, J. (2020). Towards Leveraging End-of-Life Tools as an Asset: Value Co-Creation based on Deep Learning in the Machining Industry. Hawaii International Conference on System Sciences;
Treiss, A., Walk, J., Kühl, N. (2021). An Uncertainty-Based Human-in-the-Loop System for Industrial Tool Wear Analysis. In: Dong, Y., Ifrim, G., Mladenić, D., Saunders, C., Van Hoecke, S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science, Springer.

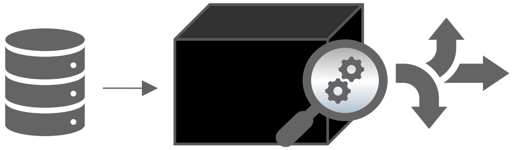
Explanations
2 dimensions help us understand the many types of explanations

…depending on the underlying ML model

- **Ante-hoc Explanation:** The “glass box” model itself is naturally interpretable (e.g., Regression, Decision Tree,…).

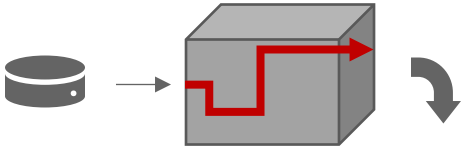


- **Post-hoc Explanation:** The “black box” model is not interpretable, and an additional interpretability method is required (e.g., Lime, SHAP, …).

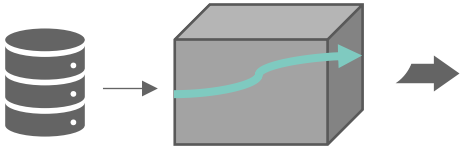


…depending on the “globality”

- **Local Explanation:** “Why does X lead to Y (in this case)?”



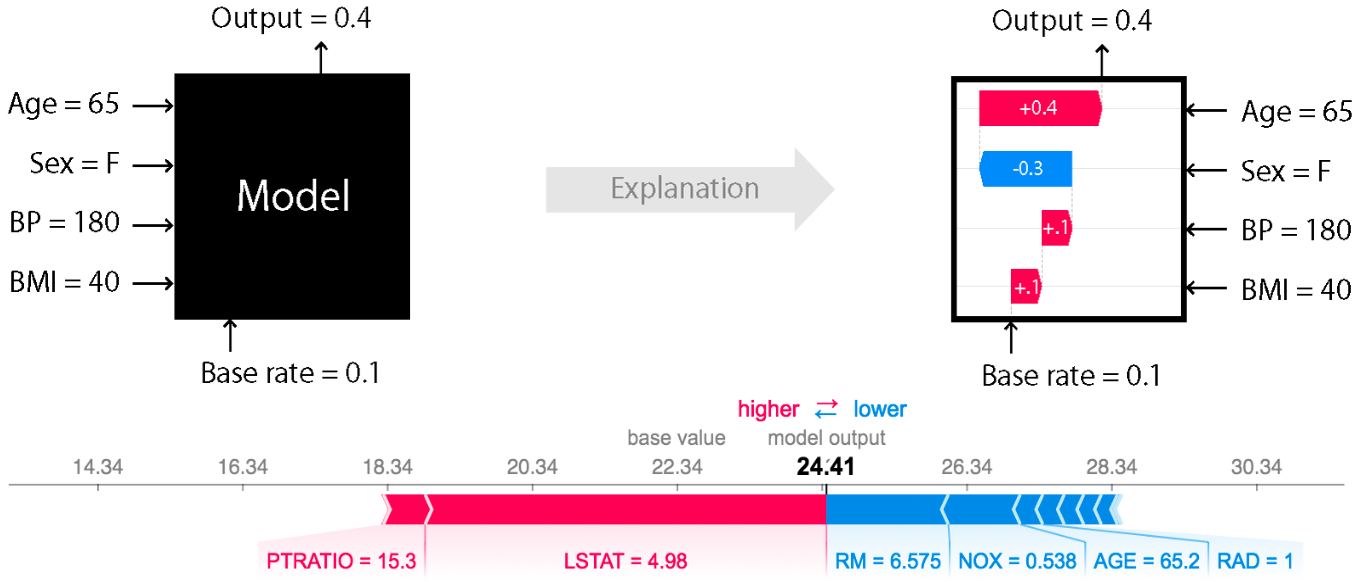
- **Global Explanation:** “How does the model work in general (e.g. on average)?”



Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.
Lundberg, S. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Explanations
SHAP is a popular framework for post-hoc explanations

“**SHAP (SHapley Additive exPlanations)** is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.”



Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.
Lundberg, S. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Explanations | Research
One question remains: are explanations always helpful?

On the one hand…

Paper 1: "Our results show that participants supported by explainable AI outperformed those supported by black-box AI because they were more likely to **follow AI predictions when they were accurate** and more likely to **overrule them when they were wrong.** [1]“

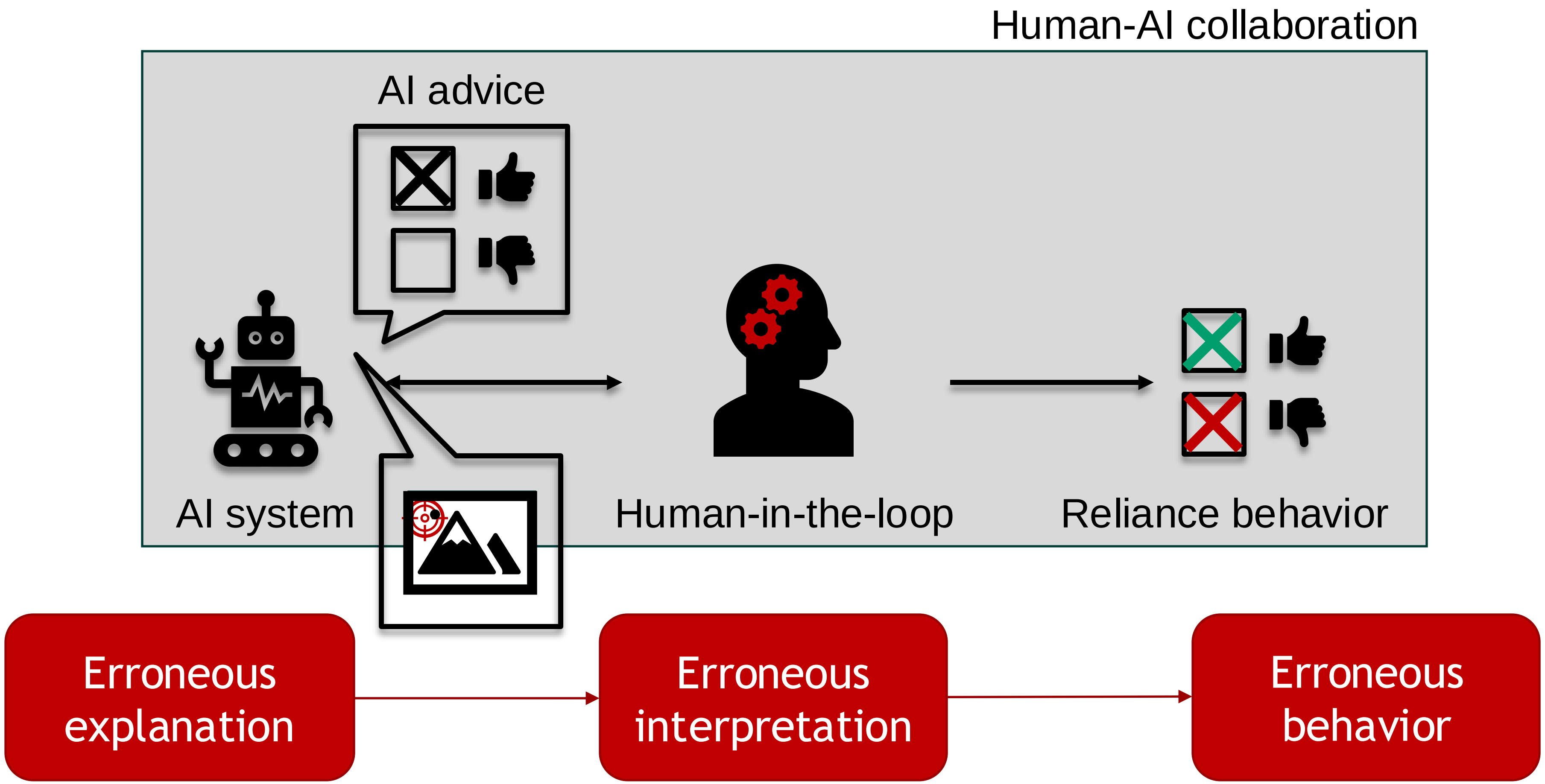
On the other hand…

Paper 2: “[…] explanations increased the chance that humans will accept the AI’s recommendation, **regardless of its correctness.**“ [2]

Current research does not agree on the helpfulness / effectiveness of explanations. Context matters!

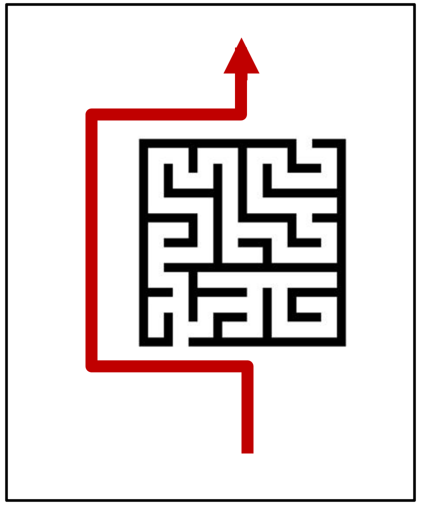
Julian Senoner, Torbjørn Netland, Stefan Feuerriegel (2021) Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing. Management Science 68(8):5704-5723. [1]
Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In Proceedings of the 2021 CHI conference on human factors in computing systems, 1-16. [2]

Explanations
Flawed explanations can have detrimental downstream impact



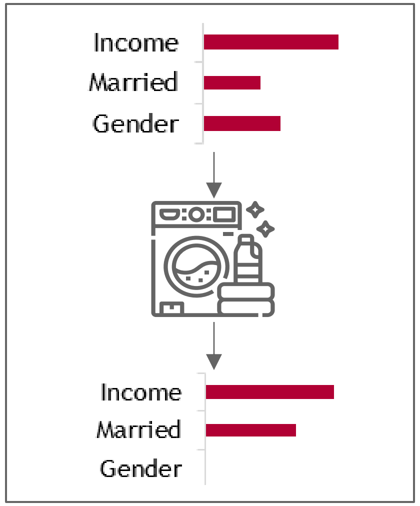
Explanations
Explanations can be straightup wrong or misleading

Reliability:
Can we trust explanations?



[1], [2]

Intentional Manipulation:
Fairwashing

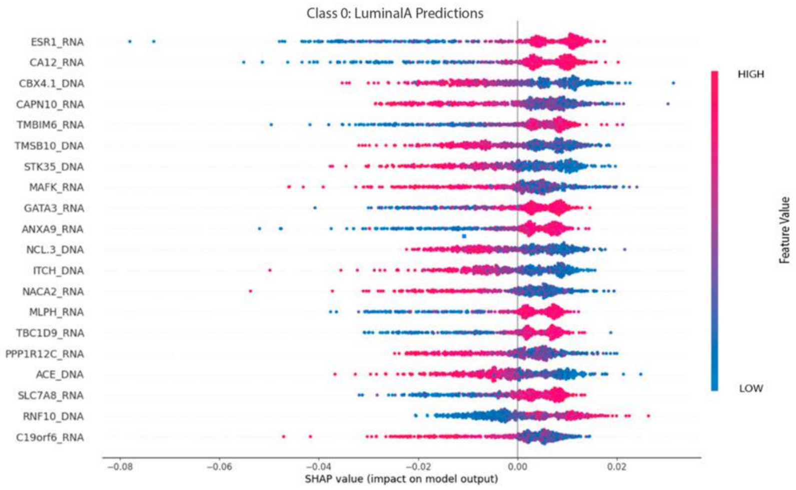


[3], [4]

Herman, B. (2017): The Promise and Peril of Human Evaluation for Model Interpretability. In: 31st Conference on Neural Information Processing Systems. [1]
Morrison, K.; Spitzer, P.; Turri, V.; Feng, M.; Kühl, N.; Perer, A. (2024): The Impact of Imperfect XAI on Human-AI Decision-Making. In: Proceedings of the ACM on Human-Computer Interaction 8(1). [2]
Aïvodji, U.; Arai, H.; Fortineau, O.; Gambs, S.; Hara, S.; Tapp, A. (2019): Fairwashing: The risk of rationalization. In: Proceedings of the 36th International Conference on Machine Learning. [3]
Le Merrer, E.; Trédan, G. (2020): Remote explainability faces the bouncer problem. In: Nature Machine Intelligence 2(9), p. 529-539. [4]

Explanations
Explanations can be overwhelming and misinterpreted…

Understandability: When information is overwhelming



[1]

Fallacies: Fairness Through Unawareness

[2]

Erroneous explanation → **Erroneous interpretation** → Erroneous behavior

Schmude, T.; Koesten, L.; Möller, T.; Tschiatschek, S. (2025): Information that matters: Exploring information needs of people affected by algorithmic decisions. In: International Journal of Human-Computer Studies 193. [1]
Deck, L.; Schoeffer, J.; De-Arteaga, M.; Kühl, N. (2024): A Critical Survey on Fairness Benefits of Explainable AI. In: ACM Conference on Fairness, Accountability, and Transparency. [2]
Images: https://github.com/shap/shap, https://imgflip.com/

Explanations
…leading to problematic behavior

Placebic Explanations:
Explanations as placebo

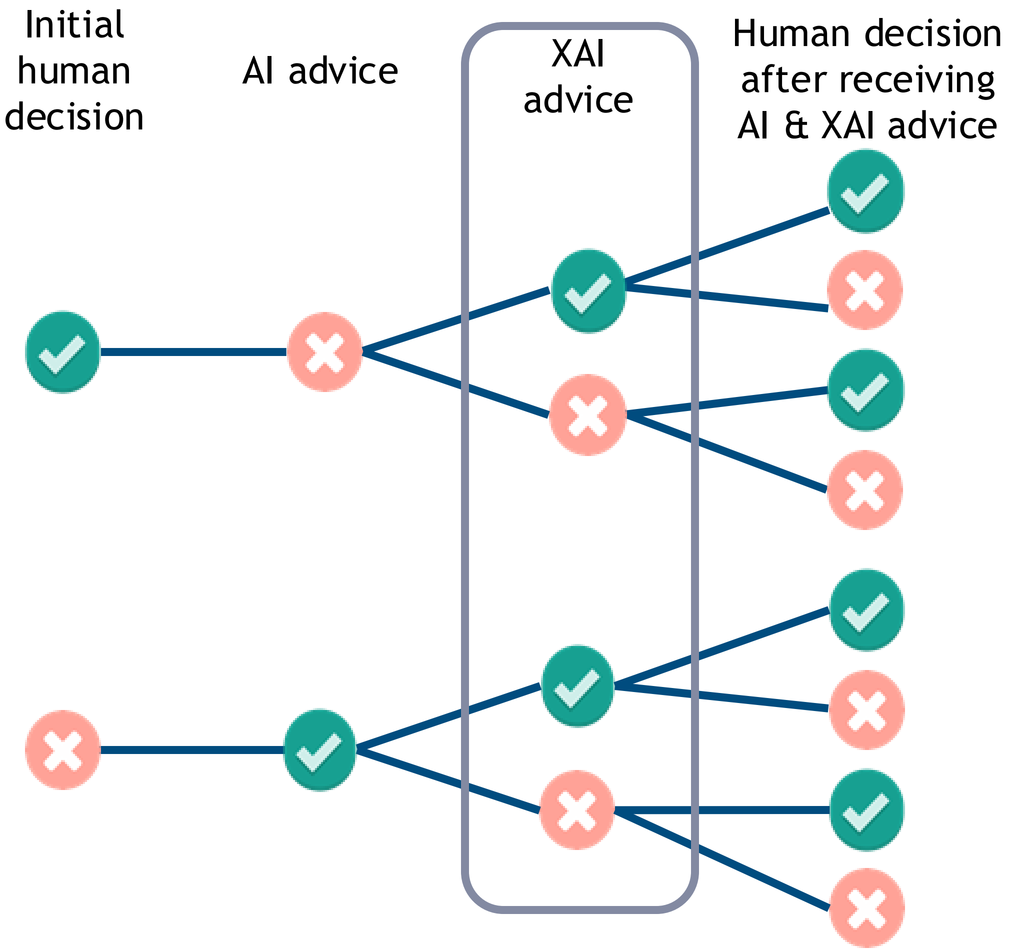
[1], [2]

Bias Alignment:
Stereotypes with explanations

[3], [4]

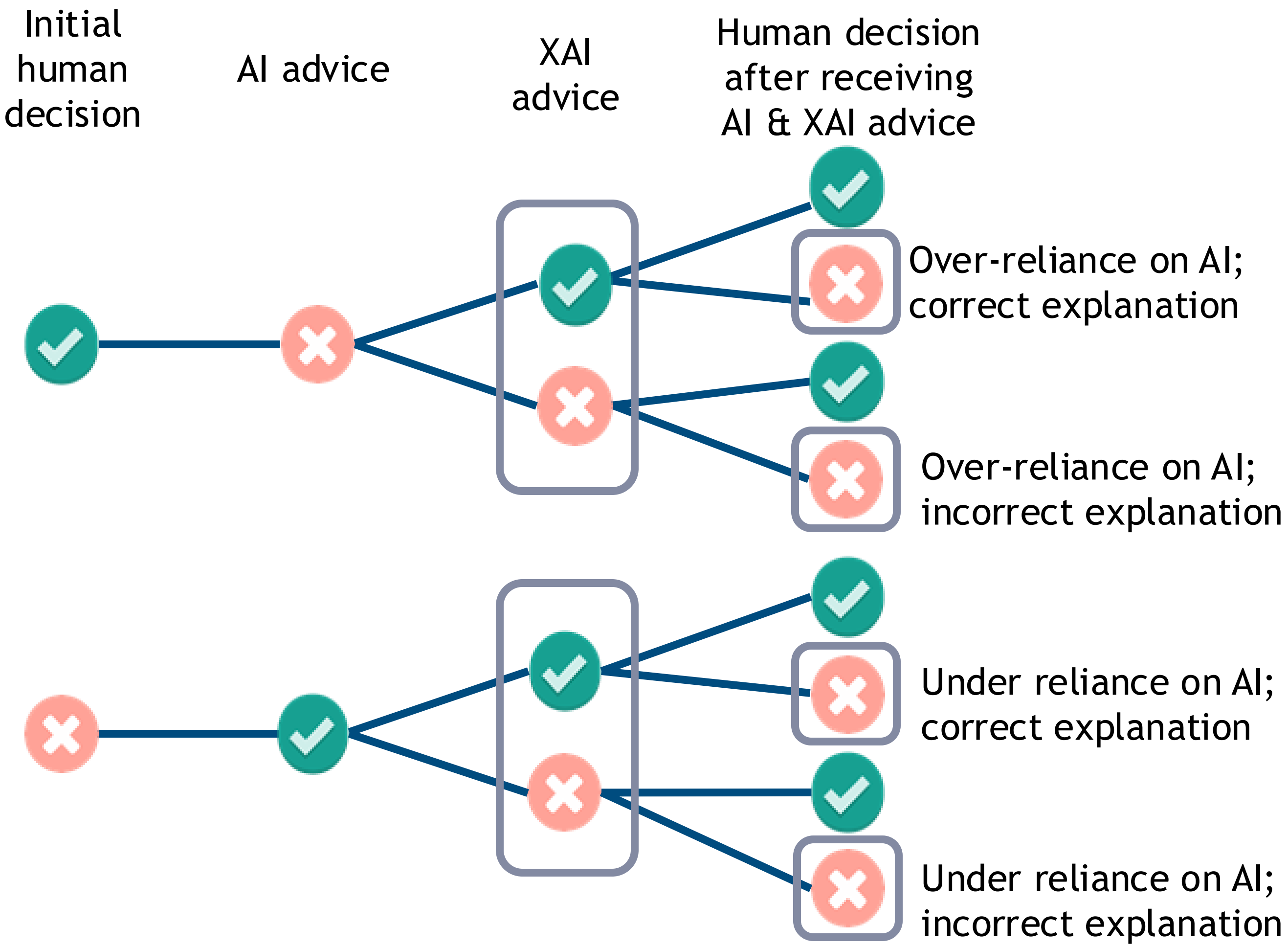
Eiband, M.; Buschek, D.; Kremer, A.; Hussmann, H. (2019): The Impact of Placebic Explanations on Trust in Intelligent Systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. [1]
Lakkaraju, H.; Bastani, O. (2020): "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. [2]
Schoeffer, J.; De-Arteaga, M.; Kuehl, N. (2024): On Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In: ACM CHI 2024. [3]
Zipperling, D., Deck, L., Lanzl, J., Kühl, N. (2024): Bias Alignment in Human-AI Teams: Effects on Decision-Making. Working paper. [4]
Images created with Midjourney

Human-AI Collaboration with Imperfect XAI
What happens when AI and XAI can both go wrong?



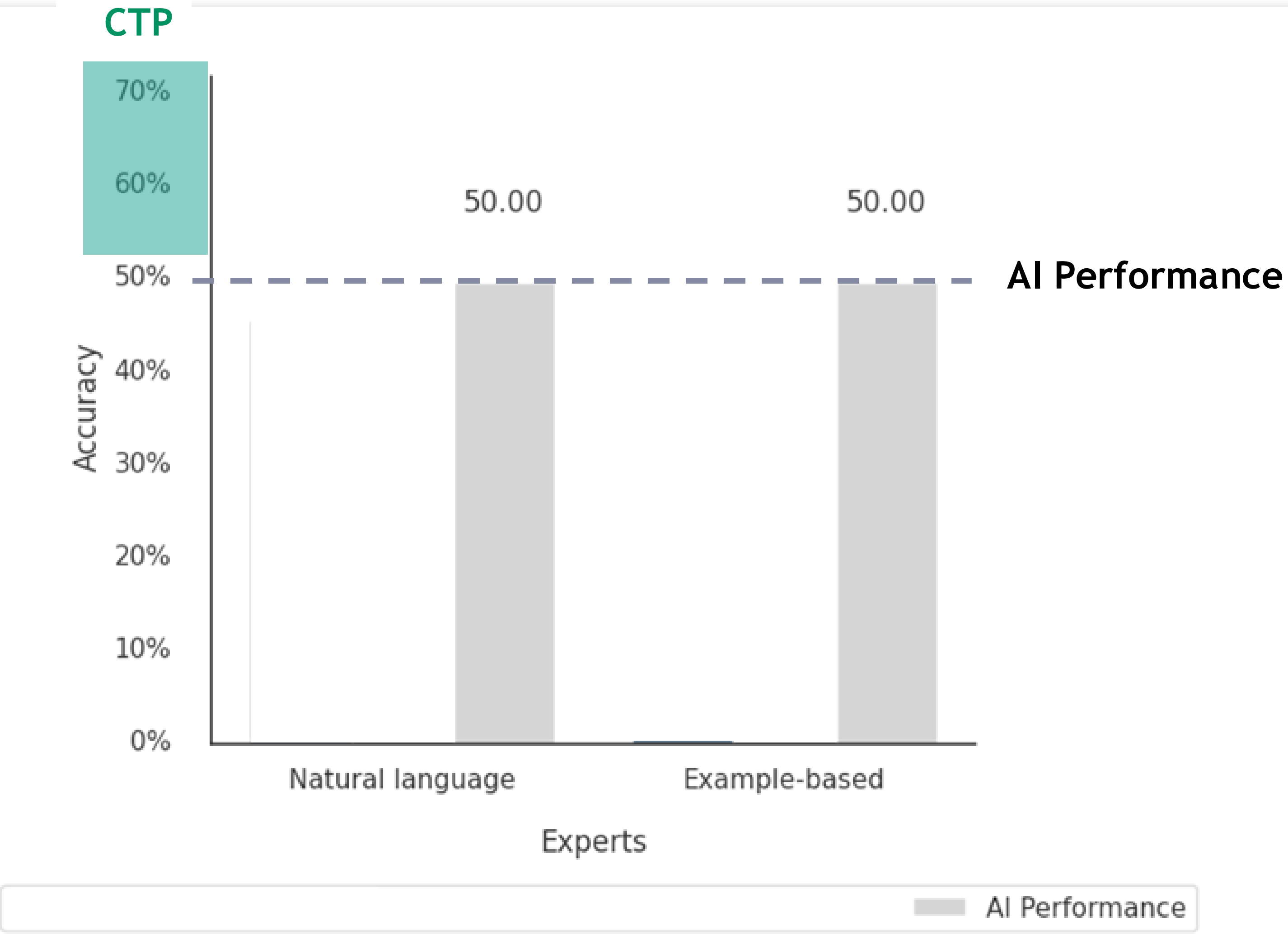
[1] Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The Impact of Imperfect XAI on Human-AI Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-39.

Human-AI Collaboration with Imperfect XAI
What happens when AI and XAI can both go wrong?



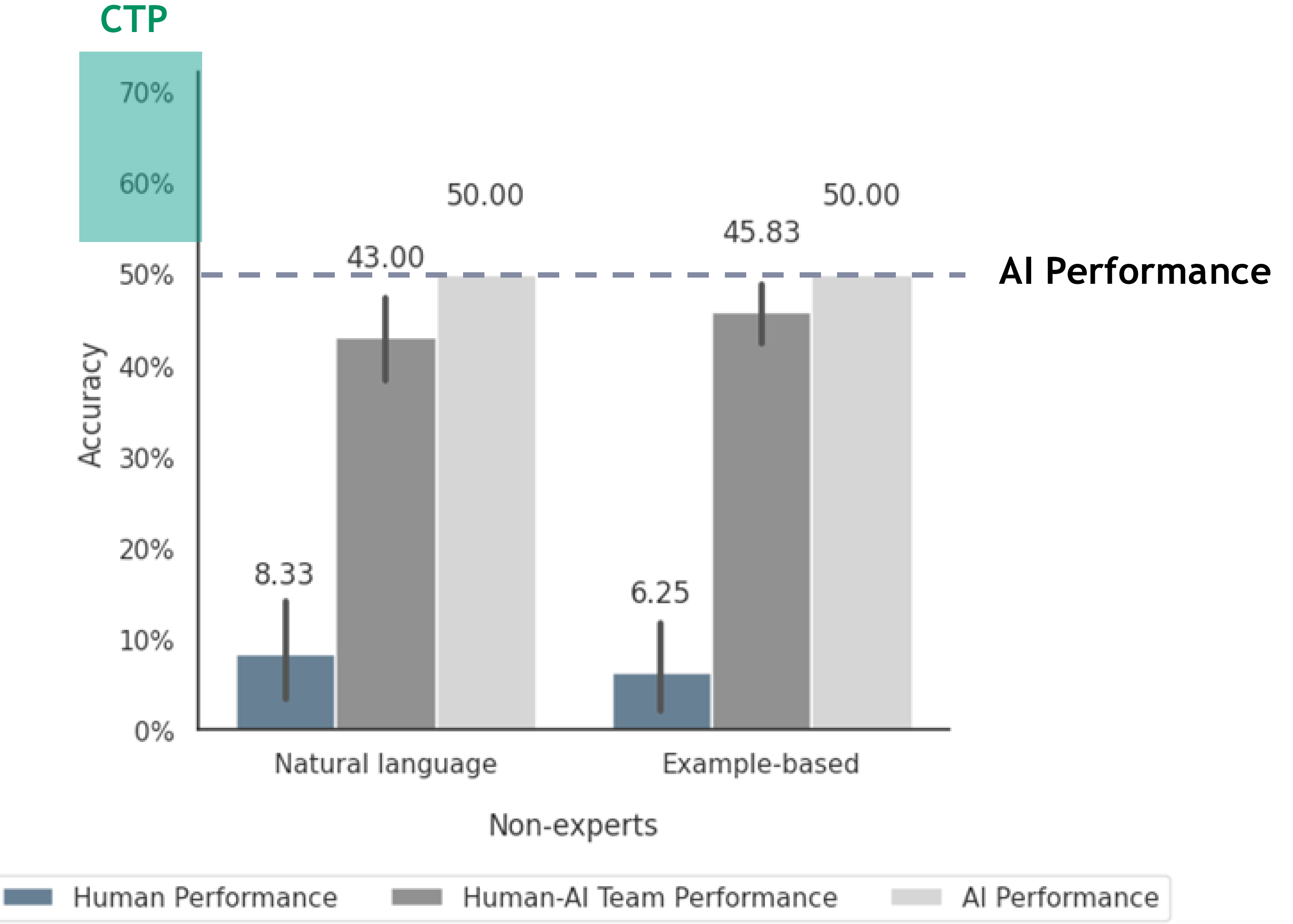
[1] Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The Impact of Imperfect XAI on Human-AI Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-39.

Complementary Team Performance (CTP) for Experts



[1] Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The Impact of Imperfect XAI on Human-AI Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-39.

Complementary Team Performance (CTP) for Non-Experts



[1] Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The Impact of Imperfect XAI on Human-AI Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-39.

Summary

Human-AI Collaboration is a viable means to leverage the complementary strengths of humans and AI. Collaboration between humans and AI consists of multiple mechanisms and complementary team performance and appropriate reliance play an important role in designing the interaction.

Collaboration between humans and AI systems is often based on mechanisms of Explainable Artificial Intelligence (XAI) and uncertainty quantifications.

“Soft factors” (e.g., perceived usefulness, trust, understanding) directly influence the adoption and use of AI systems as well as the success of Human-AI Collaboration.

Combine **complementary capabilities**

Make models **interpretable** and insights **explainable**

Validate **effectiveness** and account for **misunderstandings**

[1] Kühl (2024): Human-centric Artificial Intelligence: The road ahead