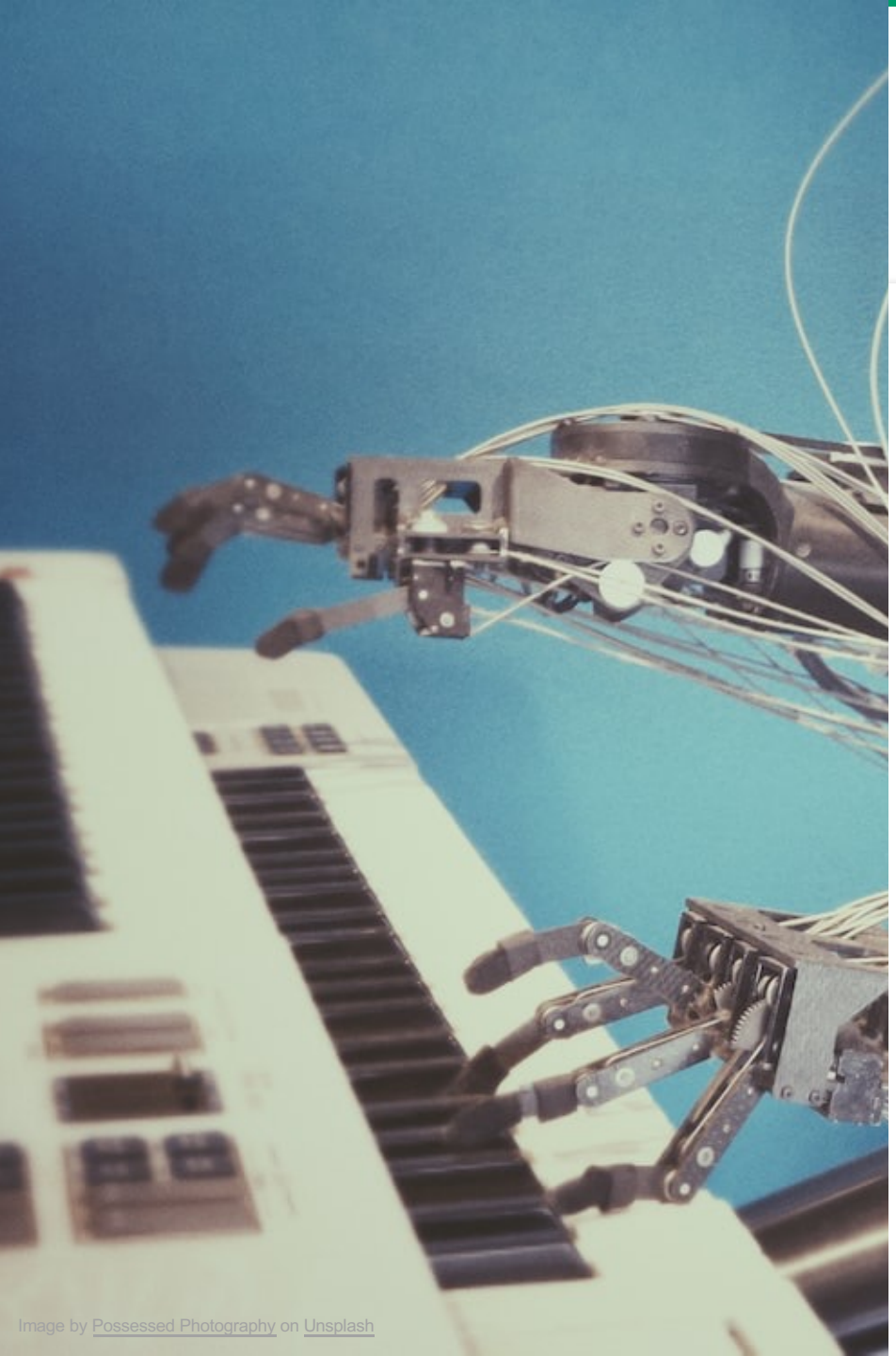


Applied Artificial Intelligence
10 - AI Ethics & Fairness

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

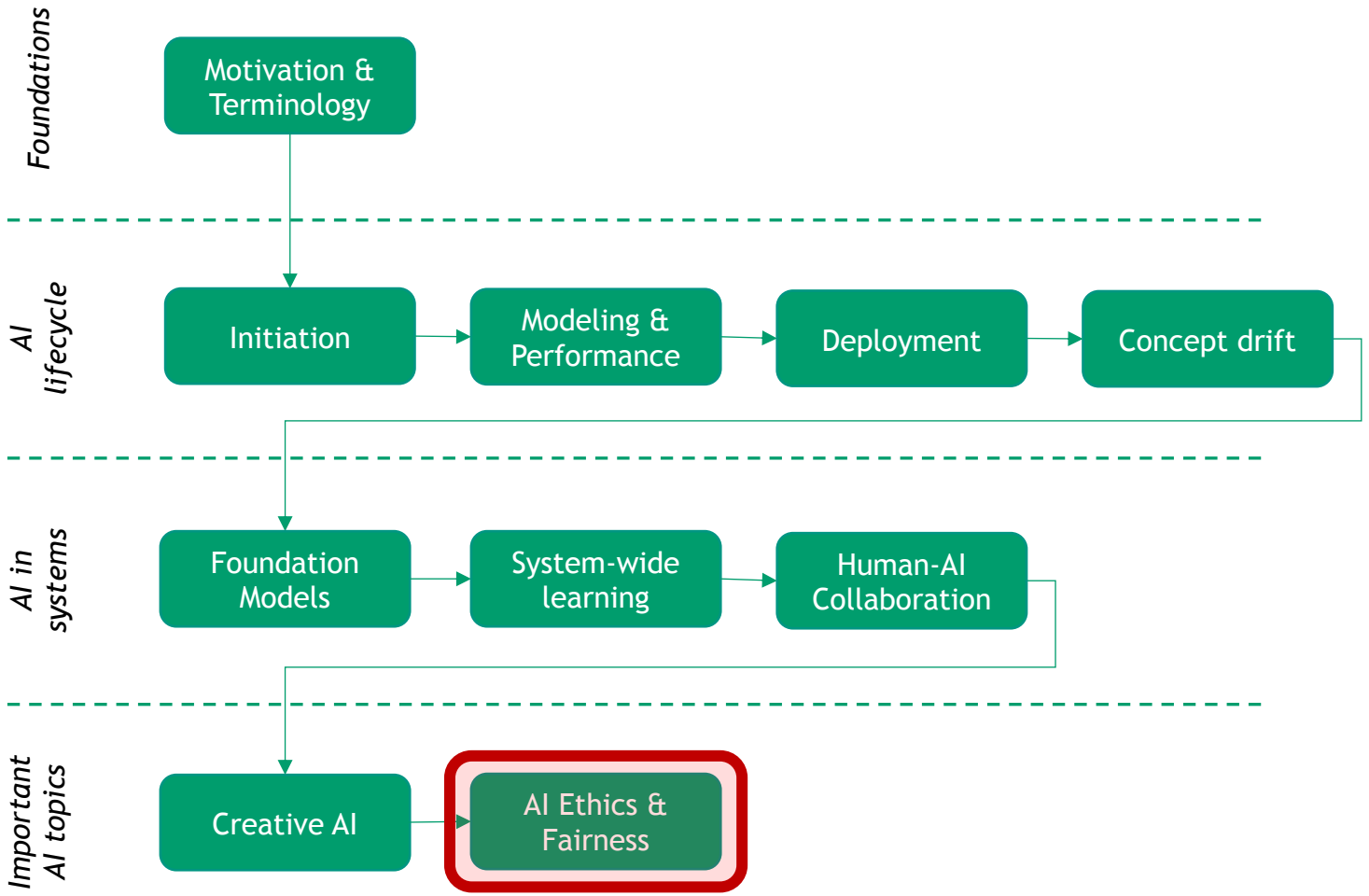
University of Bayreuth

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

Discover what future improvements in hard- and software could mean for the achievements of AI

UNDERSTAND

Understand topics related to fairness, accountability, and transparency developing ML models

INTENSIFY

Develop a critical stance towards promises and certificates of fairness

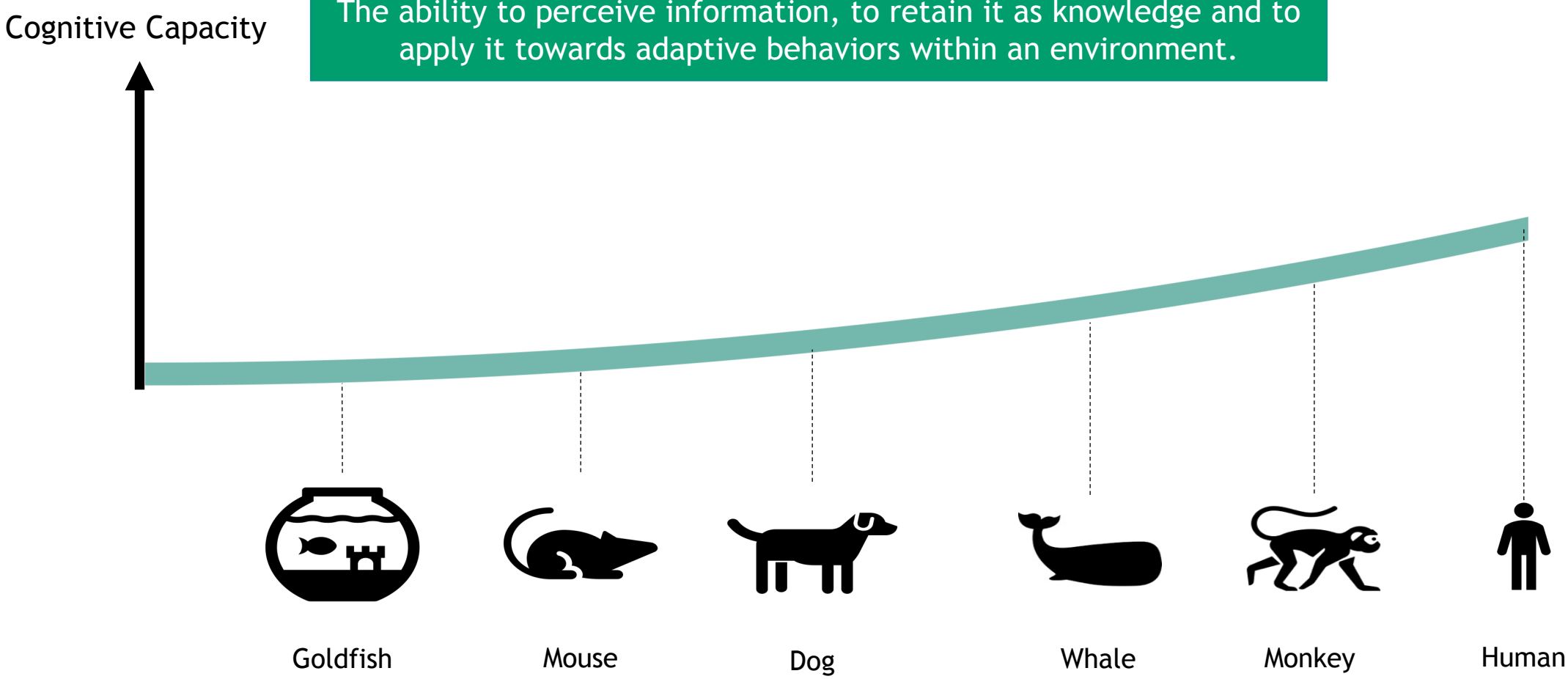
APPLY

Familiarize with basic techniques approaching unfairness in AI-based systems

1. What does the future look like?
2. The problem with accuracy
3. Can’t we just ignore sensitive information?
4. Tackling the interdisciplinary challenge of fairness
5. Outlook: We still have a long way to go

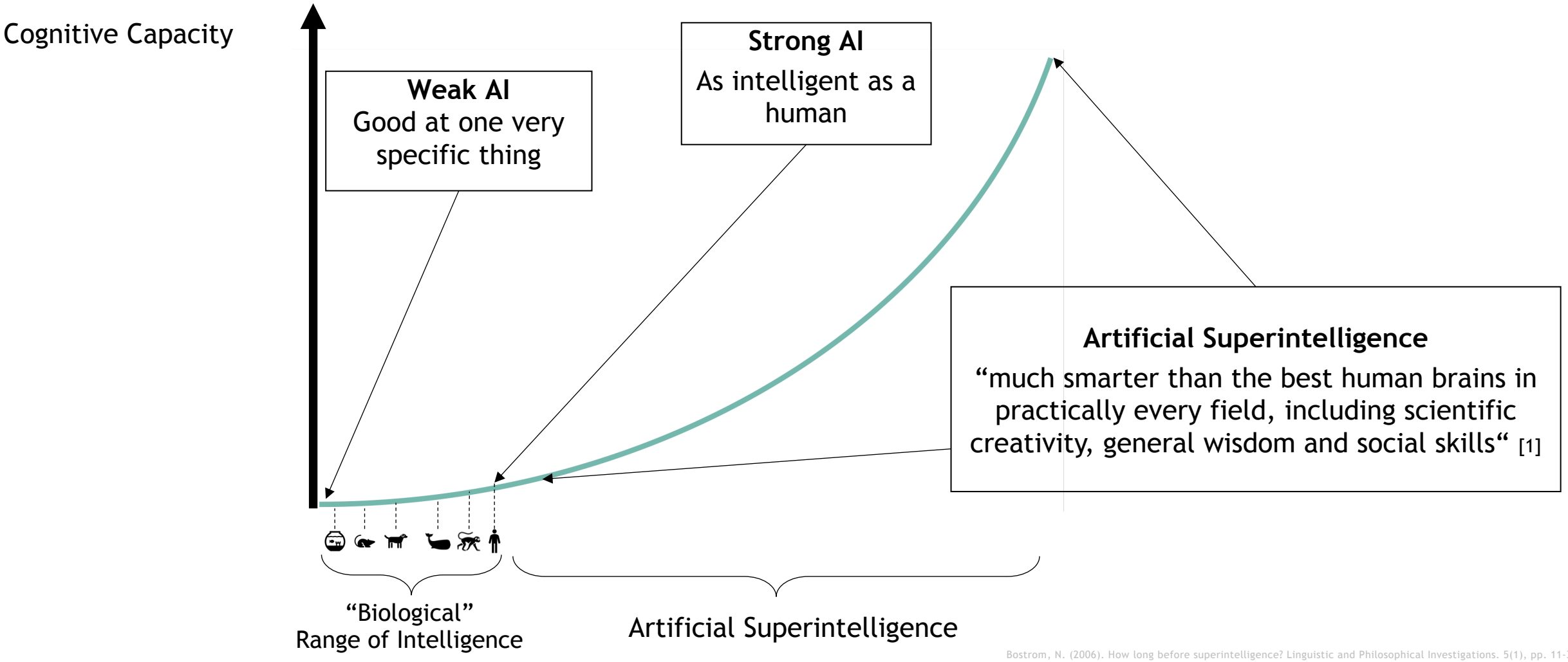
What does the future look like?
What is Intelligence?

The ability to perceive information, to retain it as knowledge and to apply it towards adaptive behaviors within an environment.



Cairo, O. (2018). External Measures of Cognition. Frontiers in Human Neuroscience 5

What does the future look like?
What is AI?



Bostrom, N. (2006). How long before superintelligence? Linguistic and Philosophical Investigations. 5(1), pp. 11-30. [1]

What does the future look like?
What is AI?

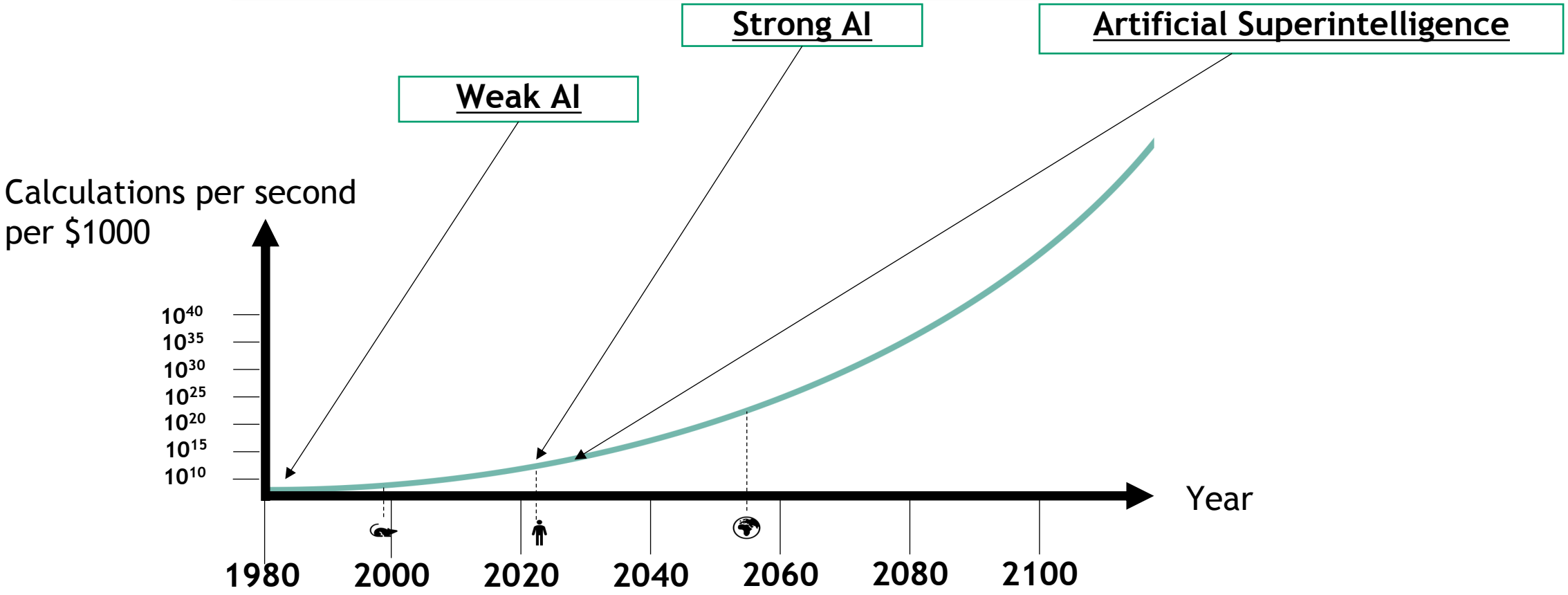
Weak AI
Good at one very specific thing

Strong AI
As intelligent as a human

Artificial Superintelligence
"much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" [1]

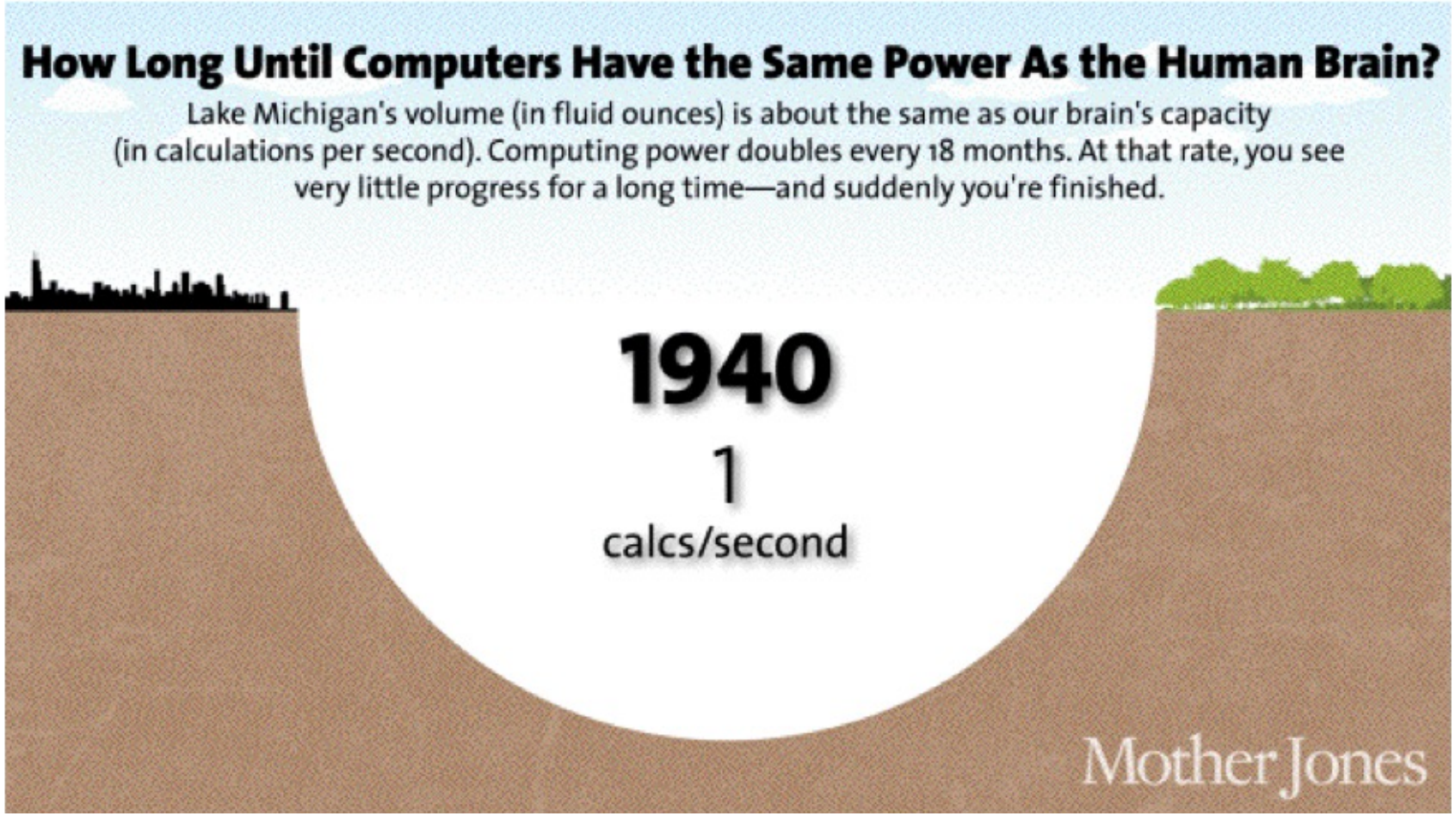
Bostrom, N. (2006). How long before superintelligence? Linguistic and Philosophical Investigations. 5(1), pp. 11-30. [1]

What does the future look like?
Projected computing power of AI



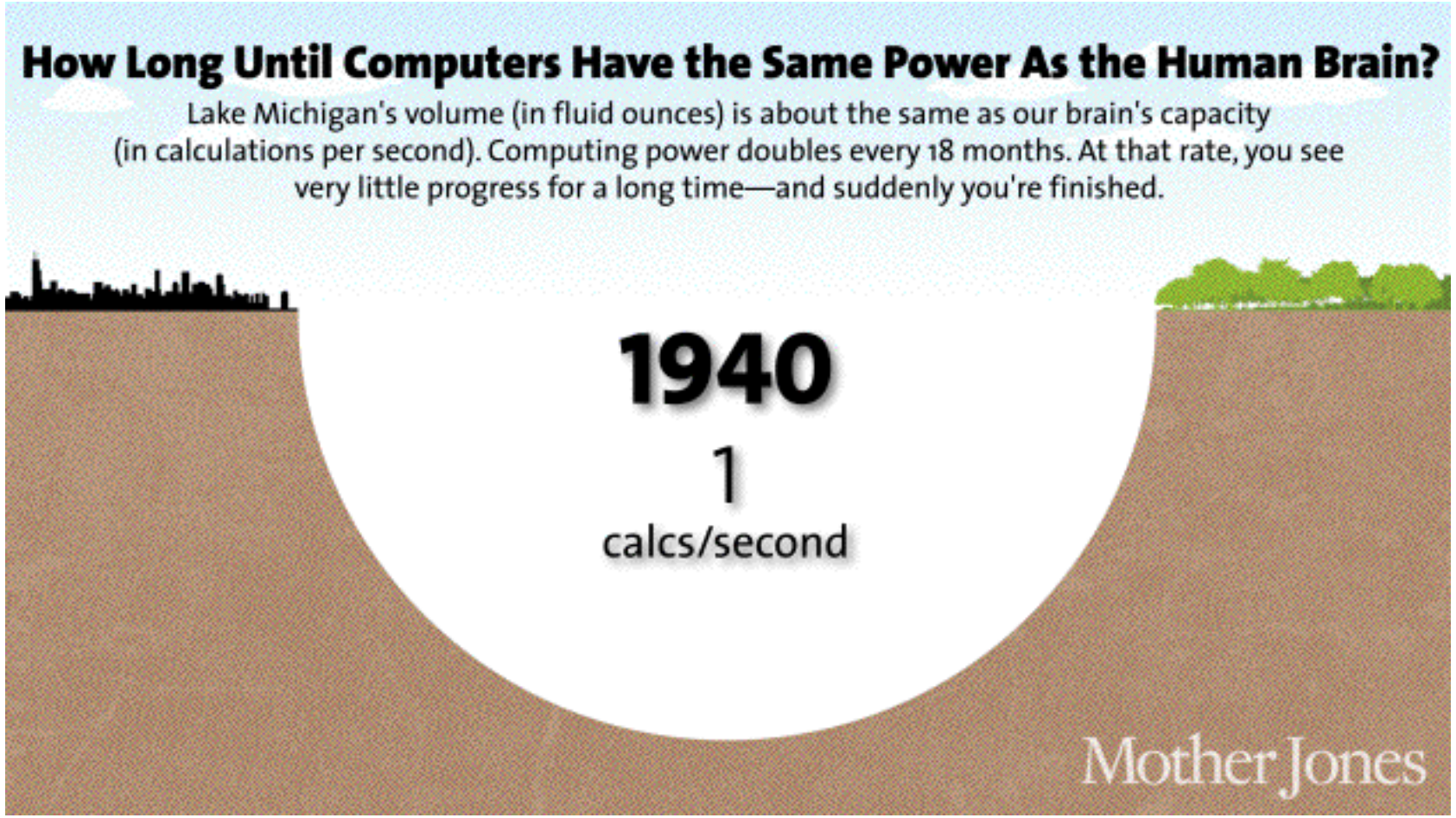
Bostrom, N. (2006). How long before superintelligence? Linguistic and Philosophical Investigations. 5(1), pp. 11-30. [1]

What does the future look like?
Projected computing power of AI



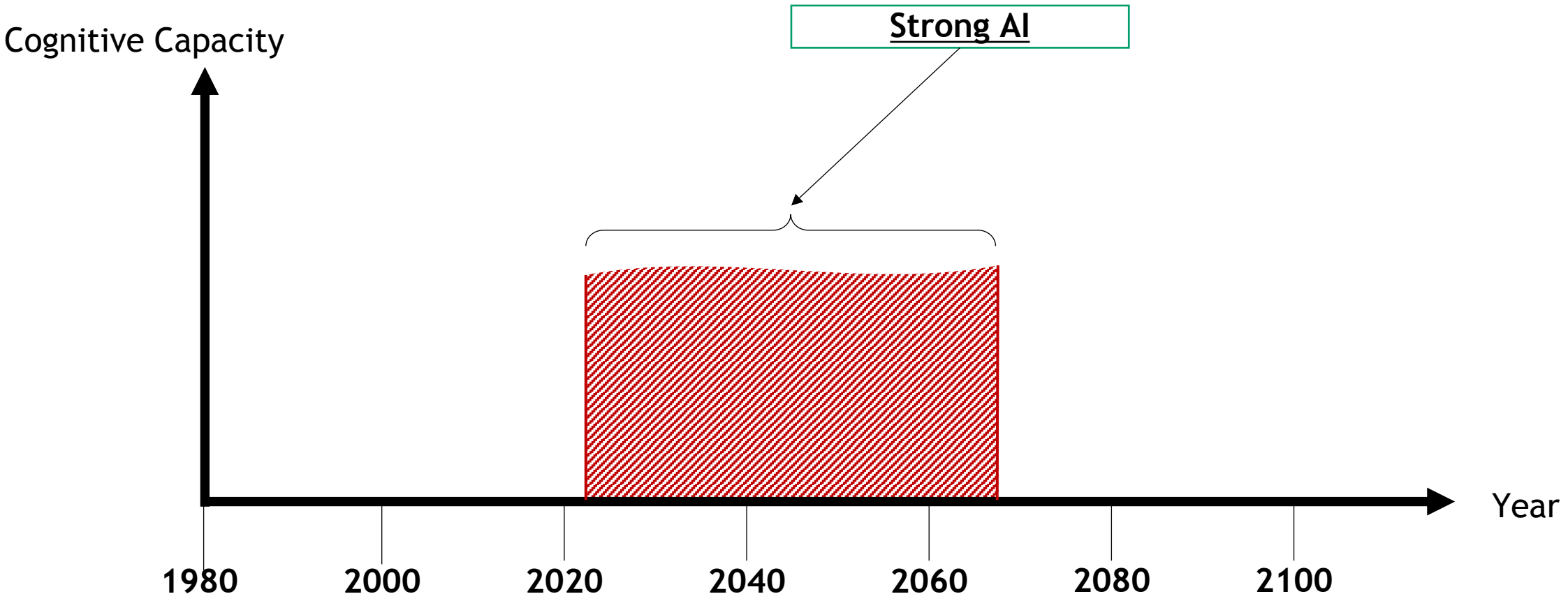
Drum, K. (2013). Welcome, Robot Overlords. Please Don't Fire Us? Mother Jones, https://www.motherjones.com/media/2013/05/robots-artificial-intelligence-jobs-automation/ (accessed Jan 07, 2025)

What does the future look like?
Projected computing power of AI



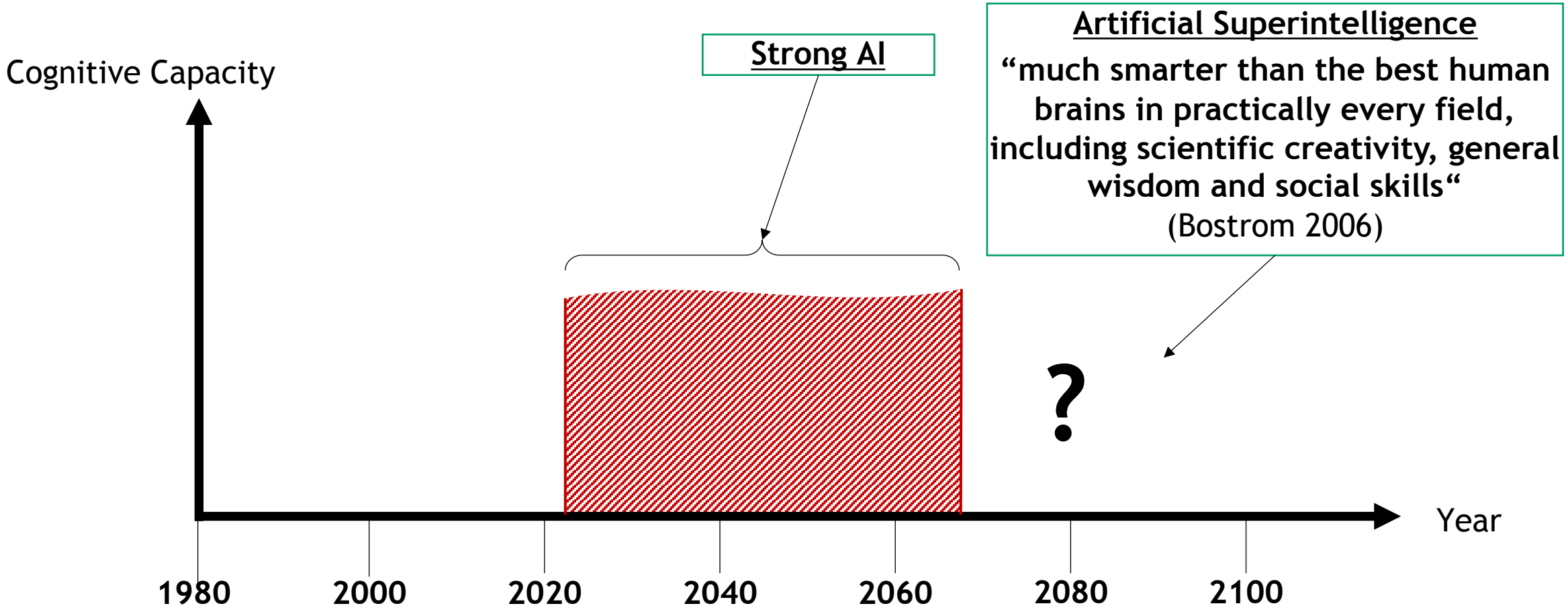
Drum, K. (2013). Welcome, Robot Overlords. Please Don't Fire Us? Mother Jones, https://www.motherjones.com/media/2013/05/robots-artificial-intelligence-jobs-automation/ (accessed Jan 07, 2025)

What does the future look like?
When will we reach Strong AI?



Dates based on Müller, V. & Bostrom, N. (2014). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. Fundamental Issues of Artificial Intelligence.

What does the future look like?
When will we reach Artificial Superintelligence?



Dates based on Müller, V. & Bostrom, N. (2014). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. Fundamental Issues of Artificial Intelligence.

What does the future look like?
What could an Artificial Superintelligence do?

Possible benefits = Positive impacts of AI on humanity

- Master nano technology
- Cure cancer
- End world hunger
- Resolve energy problems

Artificial Superintelligence

"much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" [1]

Risks = Negative impacts of AI on humanity

- Unemployment
- Population

Existential risks

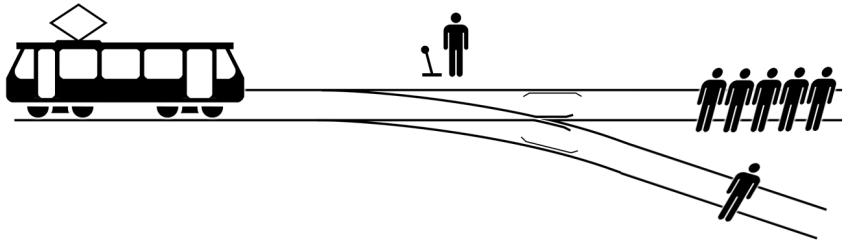
- AI self-improvements "get out of hand"
- AI alien to human morals

"There are no hard problems, only problems that are hard to a certain level of intelligence. Move the smallest bit upwards, and some problems will suddenly move from *impossible* to *obvious*. Move a substantial degree upwards, and all of them will become obvious." [2]

Bostrom, N. (2006). How long before superintelligence? Linguistic and Philosophical Investigations. 5(1), pp. 11-30. [1]
Yudkowsky, E. (1996). Starting into the singularity. [2]

What does the future look like?
Takeaways / Discussion

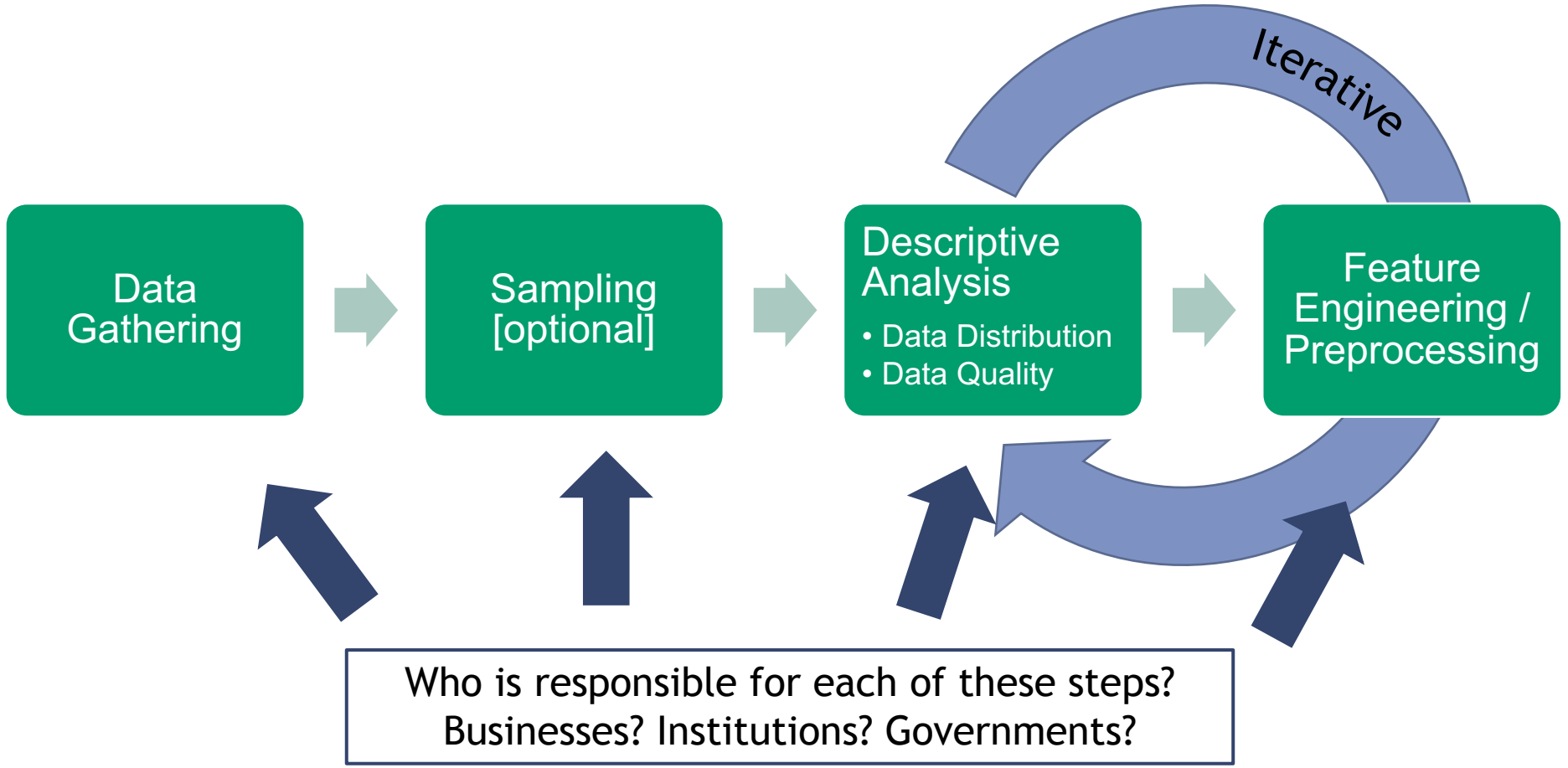
- There is **no doubt** among scientists **that** we will reach Strong AI — The only question is **when** [1]
- The arrival of Artificial Superintelligence will be **very** sudden [2]
- Artificial Superintelligence has the potential to be both very harmful and very beneficial [3]
- We need to think carefully about how we design AI [4]
  - E.g., Switchman case/trolley [5]



- Who will govern AI? ➔ "Anything that can happen will happen"

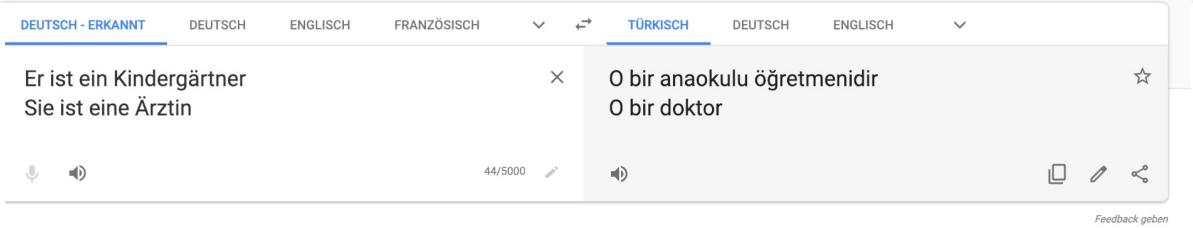
Barrat, J. (2013). Our Final Invention: Artificial Intelligence and the End of the Human Era. St. Martin's Press. [1]
Ray Kurzweil (2006). The Singularity Is Near: When Humans Transcend Biology. Penguin Books [2]
Müller, V., Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. Fundamental Issues of Artificial Intelligence, pp. 553-571. [3]
Hawking, S. et al. (2015). Open letter on artificial intelligence. [4]
Englisch, K. (1930). Untersuchung über Vorsatz und Fahrlässigkeit. [5]

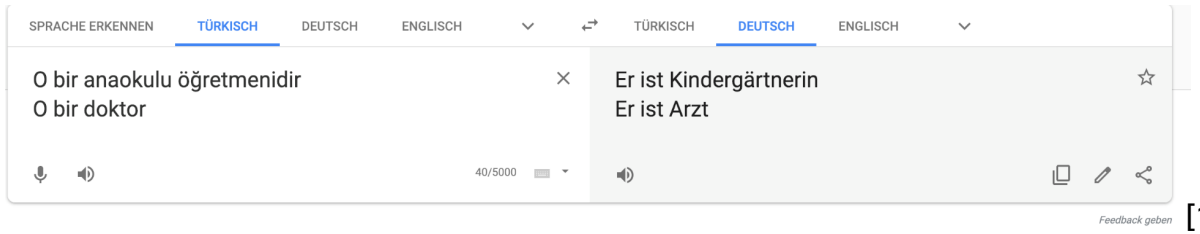
Recap second lecture: Where does AI learn from?



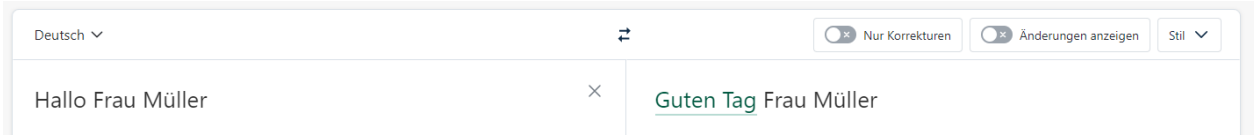
1. What does the future look like?
2. **The problem with accuracy**
3. Can’t we just ignore sensitive information?
4. Tackling the interdisciplinary challenge of fairness
5. Outlook: We still have a long way to go

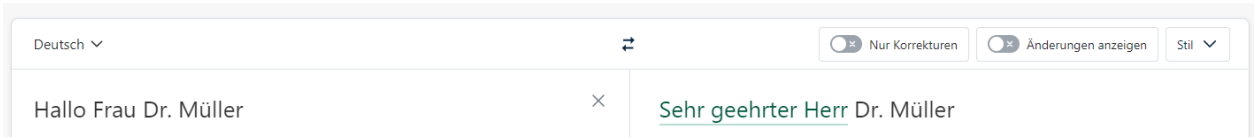
The problem with accuracy
What is happening here?





[1]





[2]

Google Translate (accessed Dec 29, 2018) [1]
DeepL (accessed Oct 26, 2023) [2]

The problem with accuracy
What is happening here?

Video: Original video from @nke_ise on X/Twitter - "Racist Soap Dispenser" (published Aug 18, 2017).

The problem with accuracy
Is “accuracy” all we should care about?

- Data scientists all over the world compete on performance benchmarks for state-of-the-art technology.
- But great accuracy does not necessarily lead to **(socially) desirable** models!
- **Garbage in, garbage out**: if the training and test data is flawed, high accuracy can be dangerous or discriminatory
- **Disparate performance**: even if the accuracy of models seems fine, it can differ unfairly between groups
- This becomes even more important as ML is used for **high-stakes decision-making.**

[1]

The New York Times

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

[2]

Kantayya, S. (2020). Coded Bias. 7th Empire Media [1]
https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html (accessed Jan 07, 2025) [2]

The problem with accuracy
ML algorithms can achieve high accuracy using “weird” information

- Task: Distinguish between **wolf and husky** pictures
- Train a classifier on set of 20 images such that **all wolf pictures have snow in the background—huskies do not**
- Model **predicts “wolf” if there is snow in the picture and “husky” otherwise**, regardless of other (seemingly) important features, still obtaining **high accuracy**!
- This is obviously a **bad model (despite high accuracy)**, but at least no one was harmed…
- Now imagine this model decides your career opportunities—and instead of snow it focuses on your skin color

(a) Husky classified as wolf (b) Explanation

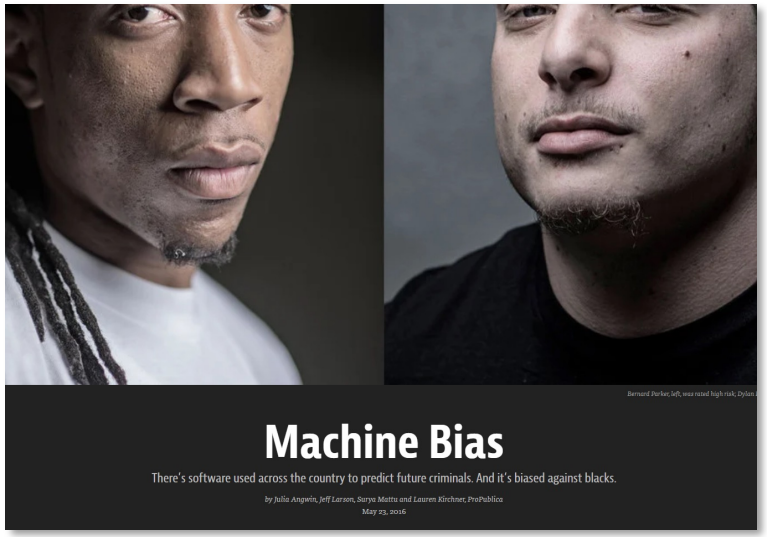
Image: Ribeiro, M., Singh, S., Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. KDD '16.

The problem with accuracy

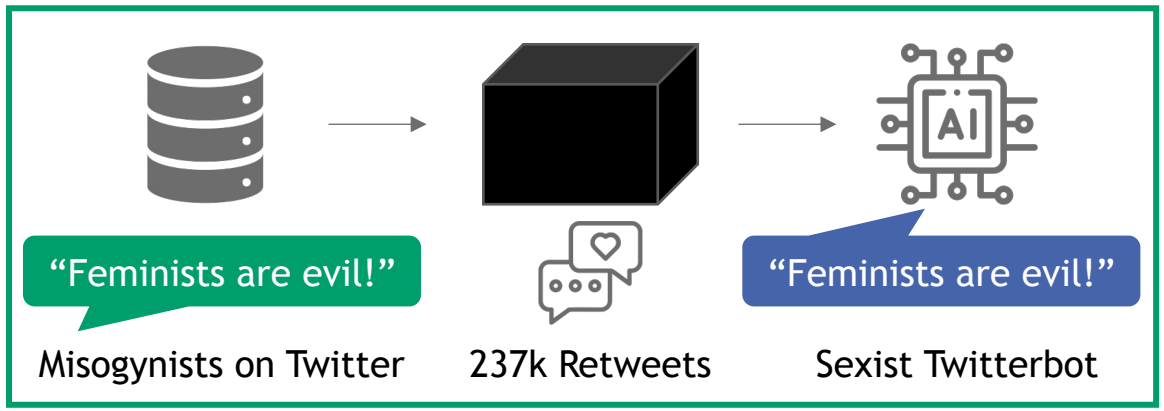
ML recognizes and reproduces learned patterns…

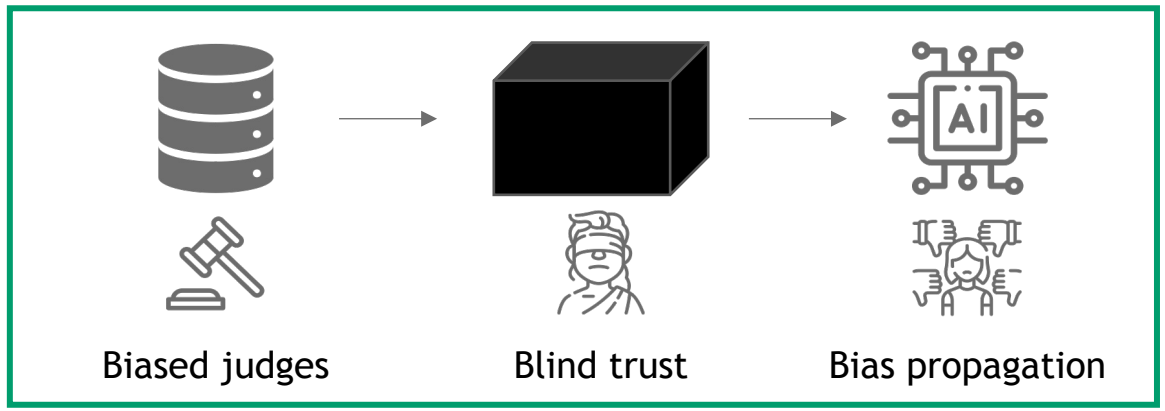
…but what if we do not approve of these patterns?

[1]



[2]





https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist (accessed Jan 07, 2025) [1]

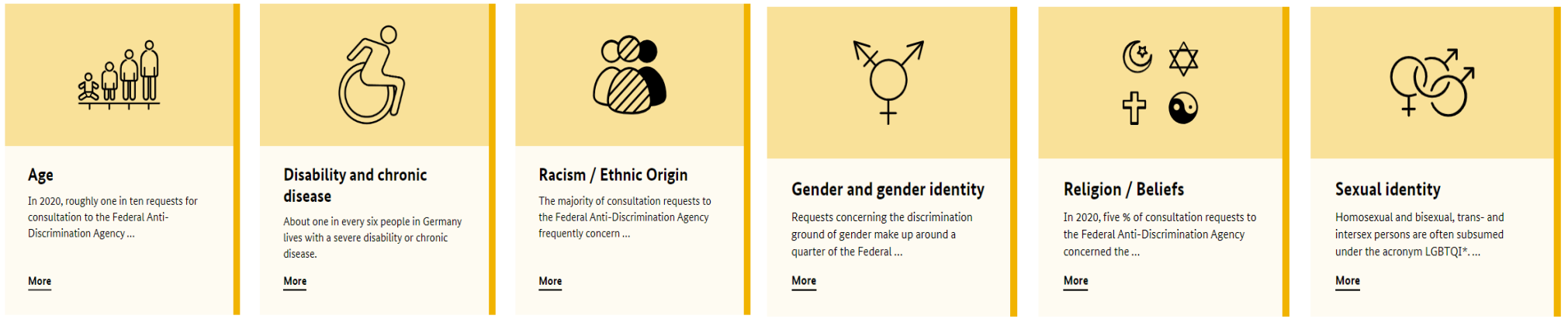
Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed Jan 07, 2025) [2]

The problem with accuracy
But is discrimination not the whole point of ML?

Discrimination
1) To be able to see the difference between two things or people.
2) To treat a person or particular group of people differently, especially in a worse way from the way in which you treat other people, because of their skin color, sex, sexuality, etc.

[1]

- ML has **no moral compass** – it will use all information available in the data to optimize the specified objective
- In supervised learning, this usually means learning labels to maximize accuracy
  - ➢ So, which information is "**legitimate**" to classify people? The AGG lists several **sensitive attributes**:



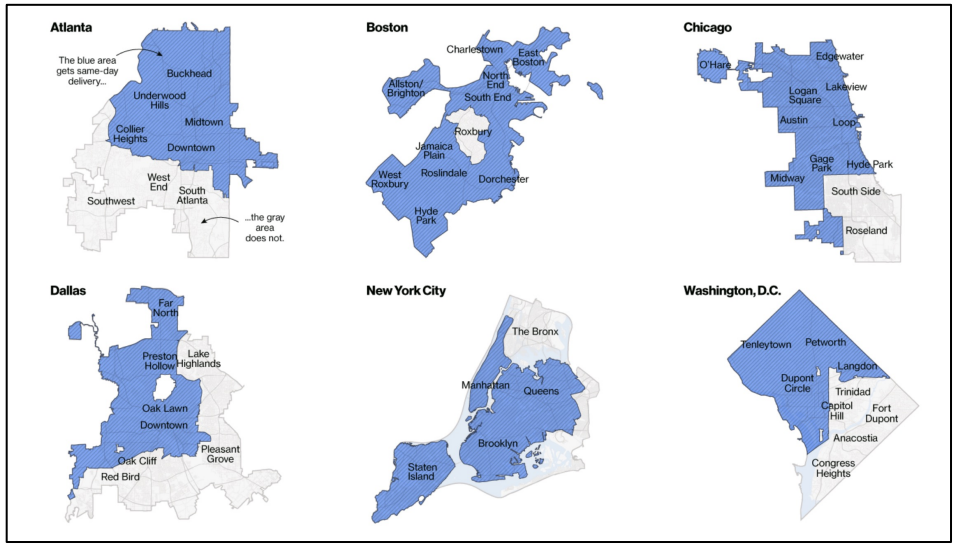
[2]

Apparently, there is "right" and "wrong" discrimination.

https://dictionary.cambridge.org/dictionary/english/discriminate (accessed Jan 07, 2025) [1]
https://www.antidiskriminierungsstelle.de (accessed Dec 07, 2024) [2]

1. What does the future look like?
2. The problem with accuracy
3. **Can’t we just ignore sensitive information?**
4. Tackling the interdisciplinary challenge of fairness
5. Outlook: We still have a long way to go

Can’t we just ignore sensitive information?



- In 2016, Amazon rolled out same-day delivery service excluding predominantly black neighborhoods.
- Amazon did **not use racial information but ZIP codes** to determine the service areas. What happened?
- Unfortunately, white neighborhoods had a substantially higher concentration of Prime members
- **Redundant Encoding**: sensitive features can be reconstructed from statistical correlations with “innocent” features
  - **Seemingly cost-efficient** and **non-discriminatory** decision-making **reinforced historical inequalities**

“Colorblindness” to sensitive information can still suffer from **redundant encoding!**

https://www.bloomberg.com/graphics/2016-amazon-same-day/ (accessed Jan 07, 2025)

Can’t we just ignore sensitive information?
Sometimes we should *proactively* include sensitive information

Can you think of examples where colorblind decisions (regarding gender or race) **do not make sense in the first place**?

Who should receive a gold medal?

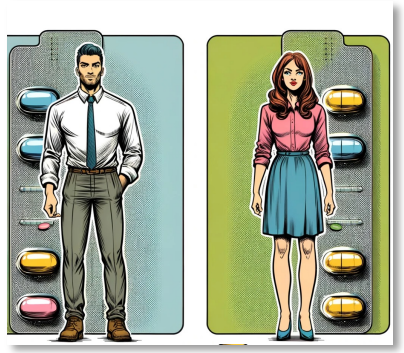
Time	Medal?
9.58s	✓
9.94s	?
10.32s	?
10.49s	?

Who should receive a gold medal now?

Time	Gender	Medal?
9.58s	m	✓
9.94s	m	✗
10.32s	m	✗
10.49s	f	✓

Who should receive iron supplementation?

Iron Level	Supplement?
5 ng/mL	✓
30 ng/mL	?
50 ng/mL	?
100 ng/mL	?



Who should receive iron supplementation now?

Iron Level	Gender	Supplement?
5 ng/mL	m	✓
30 ng/mL	m	✗
50 ng/mL	f	✓
100 ng/mL	f	✗

Sometimes **group membership** provides **meaningful information!**

Can’t we just ignore sensitive information?
Sometimes we should *proactively* include sensitive information

Can you think of examples where **preferential treatment** (as opposed to colorblindness) **is socially desired**?

Arbeitsrecht

Müssen Schwerbehinderte bevorzugt eingestellt werden? [1]

Frauenquote

Warum Frauen in Vorständen in der Minderheit bleiben [2]

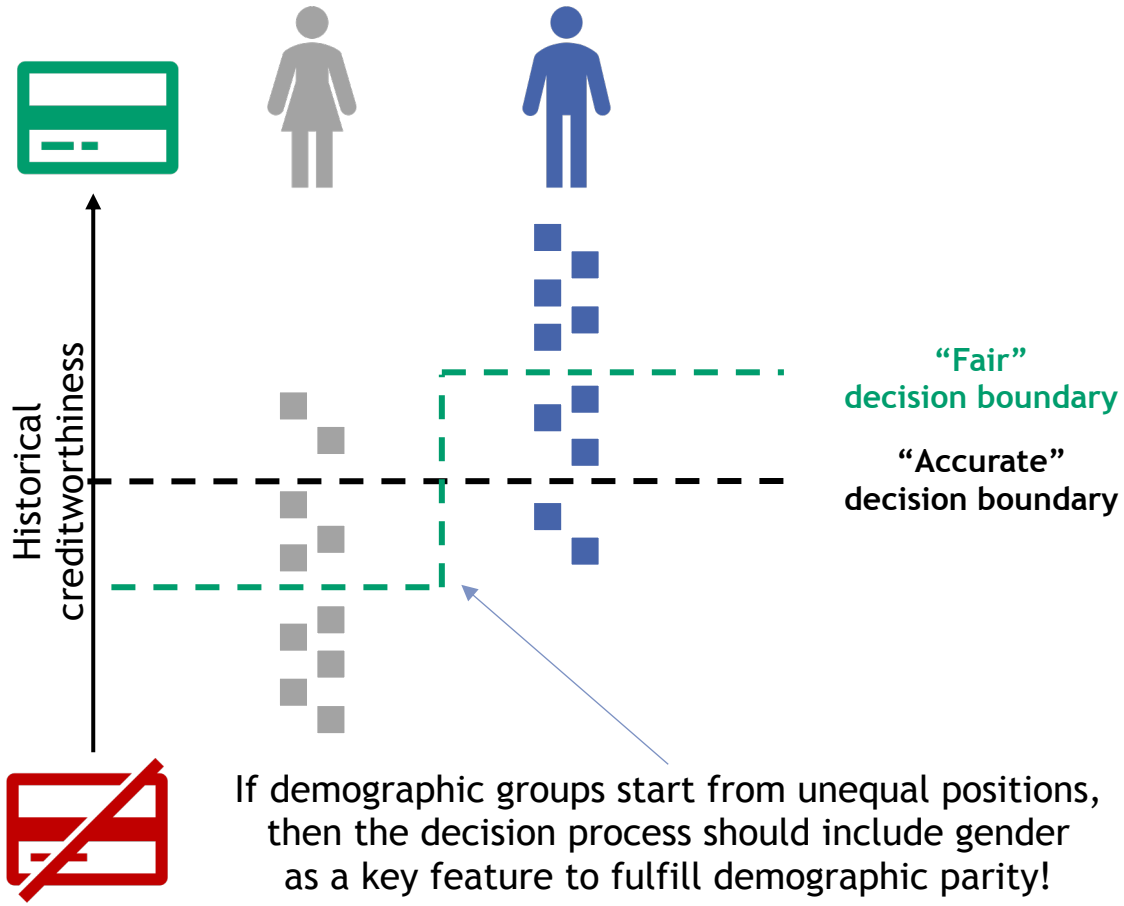
Brazil Enacts Affirmative Action Law for Universities [3]

Positive measures	“…are warranted when a group of persons is **represented far less frequently** in certain contexts than in the population at large.” [4]
Affirmative action	“…policy is not necessarily discriminatory, because its purpose and effect may be to **reduce group-based social inequality**.” [5]

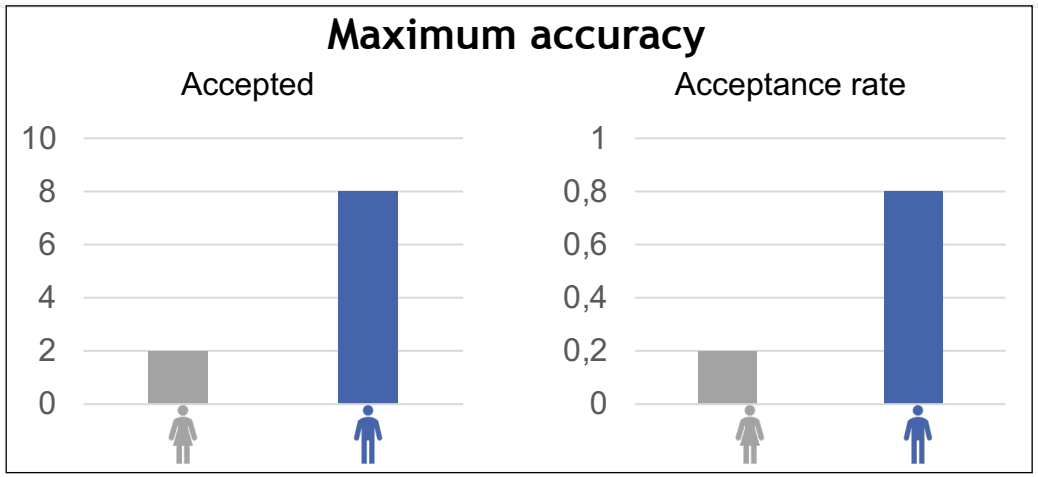
https://www.zeit.de/karriere/beruf/2012-06/arbeitsrecht-einstellung-behinderte (accessed Jan 07, 2025) [1]
https://www.handelsblatt.com/karriere/frauenquote-warum-frauen-in-vorstaenden-in-der-minderheit-bleiben/28980740.html (accessed Jan 07, 2025) [2]
https://www.nytimes.com/2012/08/31/world/americas/brazil-enacts-affirmative-action-law-for-universities.html (accessed Jan 07, 2025) [3]
Federal Anti-Discrimination Agency (2019). Guide to the General Equal Treatment Act [4]
Nyarko, J.; Goel, S.; Sommers, R. (2022). Breaking Taboos in Fair Machine Learning: An Experimental Study. In: Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO ’21 [5]

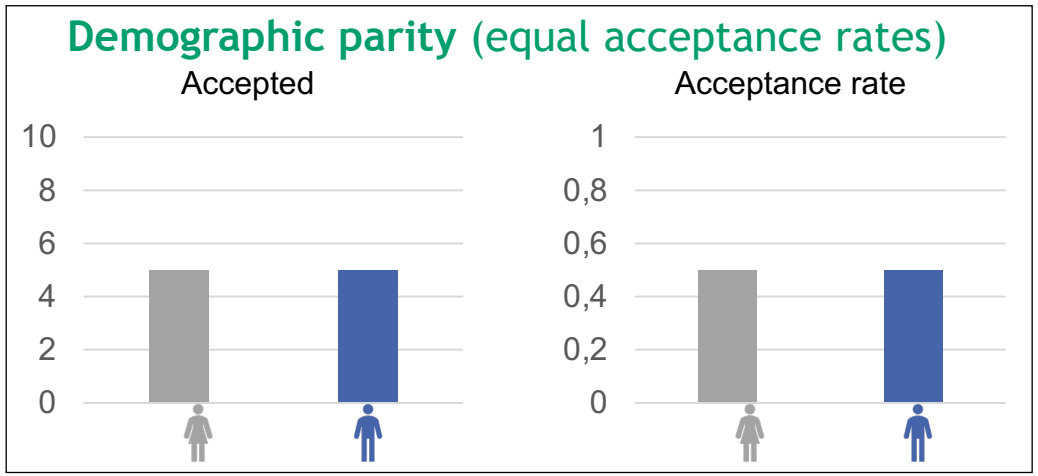
Can’t we just ignore sensitive information?
“Fair affirmative action”: A “*fair*” outcome can rely on an “*unfair*” process

Task: Assign 10 credits *fairly* among the following applicants.



If demographic groups start from unequal positions, then the decision process should include gender as a key feature to fulfill demographic parity!





Dwork, C. et al. (2012). Fairness through awareness. ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214-226.

1. What does the future look like?
2. The problem with accuracy
3. Can’t we just ignore sensitive information?
4. **Tackling the interdisciplinary challenge of fairness**
5. Outlook: We still have a long way to go

Tackling the interdisciplinary challenge of fairness
First, we need to define what *fairness* means

AI Fairness is an **emergent field of research** that uniquely unifies **multiple domains** with **varying definitions.**

[1]

Fairness notions are **multidimensional** and often **conflicting!**

Image created with Word-Cloud-Tool.

Mulligan, D., Kroll, J., Kohli, N., & Wong, R. (2019). This Thing Called Fairness. Proceedings of the ACM on Human-Computer Interaction [1]

Tackling the interdisciplinary challenge of fairness
Why not stick to non-discrimination law?

Own Research

AI Fairness vs. Non-Discrimination

- Current non-discrimination regulation (e.g., AGG) is (primarily) designed for human decision-making
- Current digital regulation (e.g., GDPR) is not (primarily) designed for non-discrimination
- Fairness is a much broader concept than discrimination [2]
  - subjective notion vs. explicit law
  - pursuing equality vs. avoiding harm
  - prescriptive approach vs. reactive measures

General Data Privacy Regulation (GDPR)

In order to ensure fair and transparent processing […] the data controller […] should prevent […] discriminatory effects on natural persons on the basis of

sensitive features
- racial or ethnic origin,
- political opinion,
- religion or beliefs,
- trade union membership,
- genetic or health status
- or sexual orientation […]

[1]

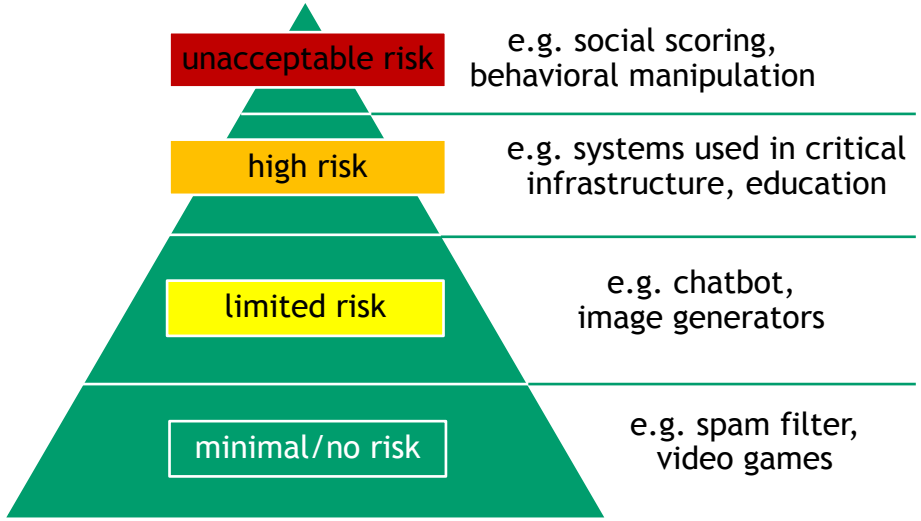
⚠ Algorithmic discrimination is often **without "human" intent** and **difficult to prove.**

GDPR (Recital 71) [1]
Deck, L., Müller, J., Braun, C., Zipperling, D. (2024). Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness. European Workshop on Algorithmic Fairness [2]

Tackling the interdisciplinary challenge of fairness
Does the AI Act tell us what fairness is?

EU AI Act
- Most ambitious legislation on AI, could become global standard
- Stepwise entering into force starting from August 2024
- High-risk regulation entering into force in August 2026

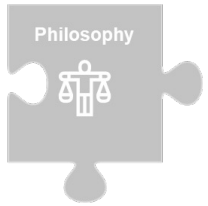
Why should we care?
- Philosophers, psychologists, computer scientists can have opinions on what fairness *should* be
- However, in the end, the law needs to *decide* a given case
- These cases are in turn guided by interdisciplinary discourse



> **Training, validation and testing data sets** shall be subject to **data governance and management practices** appropriate for the intended purpose of the AI system […] in view of **possible biases** that are likely to […] lead to **discrimination prohibited under Union law** […]. [1]

The AI Act does not tell us what fairness is, but it underlines responsibilities for providers and deployers. Future case law will set the boundaries for acceptable industry practices.

AI Act, Article 10(2)f [1]



Tackling the interdisciplinary challenge of fairness
What makes discrimination “wrong”?

Taste-based Discrimination (with discriminatory intent)

- “Differential treatment on the basis of membership of a salient social group–e.g. gender or ‘race’–by those with decision-making power to distribute harms or benefits.”

➢ Who has intentions in AI settings? The algorithm? The human-in-the-loop? The designers? The provider?

Statistical Discrimination (not necessarily intended)

- “The use of statistical generalisations about groups to infer attributes or future behaviours of members of those groups.”

➢ AI fails to treat people as individuals by design! Are generalisations acceptable when they are sufficiently precise?

[1]

Discriminatory club bouncers

Allocational harms

- Allocating resources unevenly between groups, rejecting access to opportunities

Representational harms

- Representing groups in an unfavorable way or ignoring their existence in the first place

[2]

Negative emotions in Midjourney

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of Machine Learning Research, pp. 1-11. [1]
Crawford, K. (NIPS 2017). https://www.youtube.com/watch?v=fMym_BKWQzk (accessed Jan 07, 2025) [2]

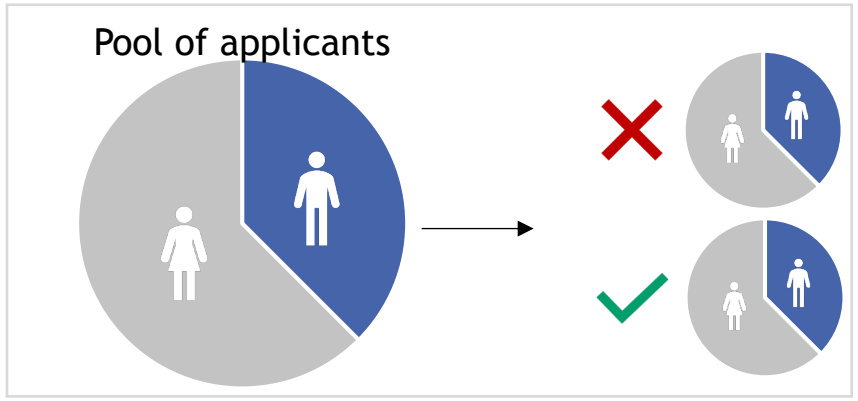
Tackling the interdisciplinary challenge of fairness
Formal fairness metrics: Demographic parity

Demographic parity: "Acceptance rate is equal for all sensitive groups"

$$\Pr(\hat{Y} = 1 | A = a) = \Pr(\hat{Y} = 1 | A = b)$$

- Example: **60% females and 40% males apply** to university

➢ The pool of admitted students should include **60% females and 40% males**



Variable	Description
X	Features
A	Sensitive Feature(s)
Y ∈ {0,1}	"Ground Truth"
Ŷ ∈ {0,1}	Prediction/Decision

		Ground Truth Y: Positive	Ground Truth Y: Negative
Predictions Ŷ	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

➢ Does this rule out unfair practices?

⚠ No information on accuracy → possibly arbitrary outcomes

⚠ Differing decision thresholds might violate individual fairness

Barocas, S. Hardt, M., Narayanan, A (2019). Fairness and Machine Learning, fairmlbook.org (accessed Jan 07, 2025)

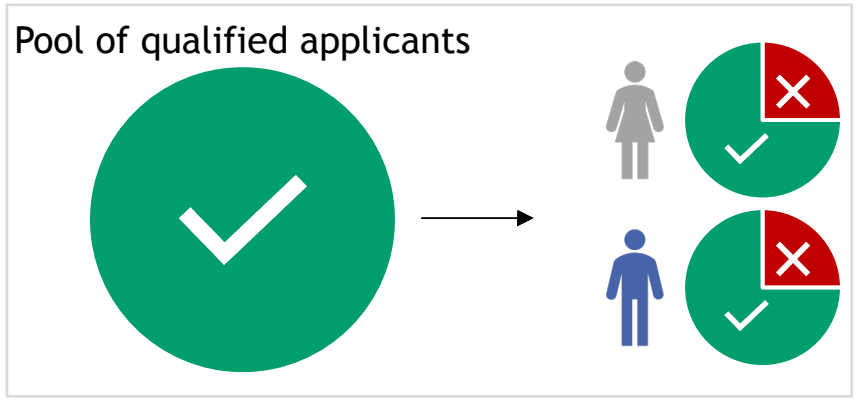
Tackling the interdisciplinary challenge of fairness
Formal fairness metrics: Equal opportunity

Equal opportunity: "True positive rates are equal for all sensitive groups"

$$\Pr(\hat{Y} = 1 | Y = 1, A = a) = \Pr(\hat{Y} = 1 | Y = 1, A = b)$$

- Example: Of **20 qualified female applicants**, we **correctly accepted 15**

➢ The **true positive rate of ¾** should also apply to male applicants



Variable	Description
X	Features
A	Sensitive Feature(s)
$Y \in \{0,1\}$	"Ground Truth"
$\hat{Y} \in \{0,1\}$	Prediction/Decision

Predictions $\hat{Y}$ \ Ground Truth Y	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

➢ Would it make sense to equalize **error** rates?

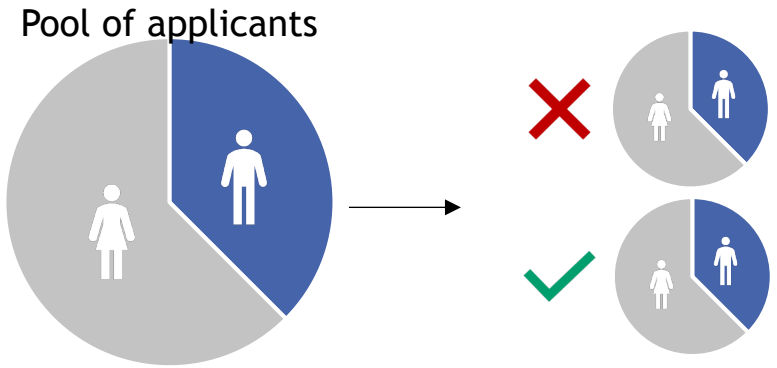
⚠ Is it fair to sacrifice accuracy in the privileged group?

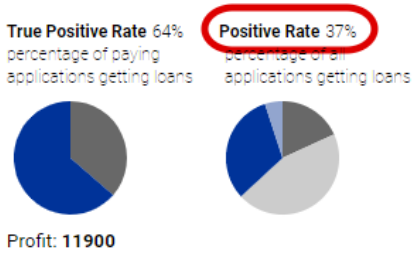
⚠ Or should we rather improve accuracy for the unprivileged group?

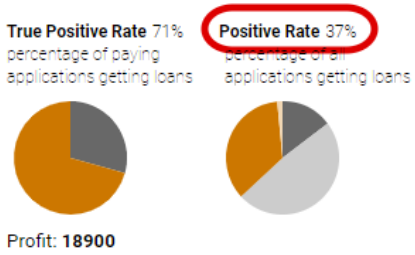
Barocas, S. Hardt, M., Narayanan, A (2019). Fairness and Machine Learning, fairmlbook.org (accessed Jan 07, 2025)

Tackling the interdisciplinary challenge of fairness
Fairness metrics are often conflicting and have varying results!

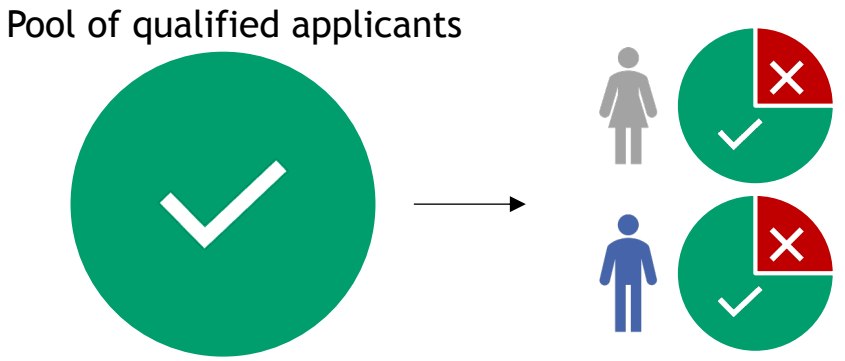
Statistical Parity

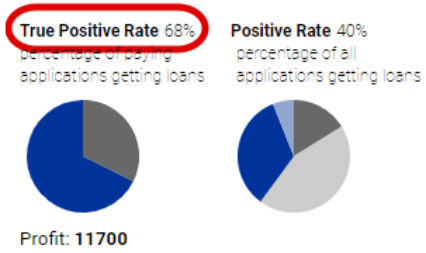


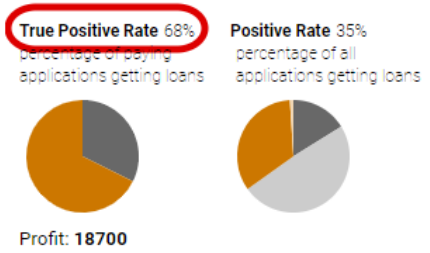




Equal Opportunity





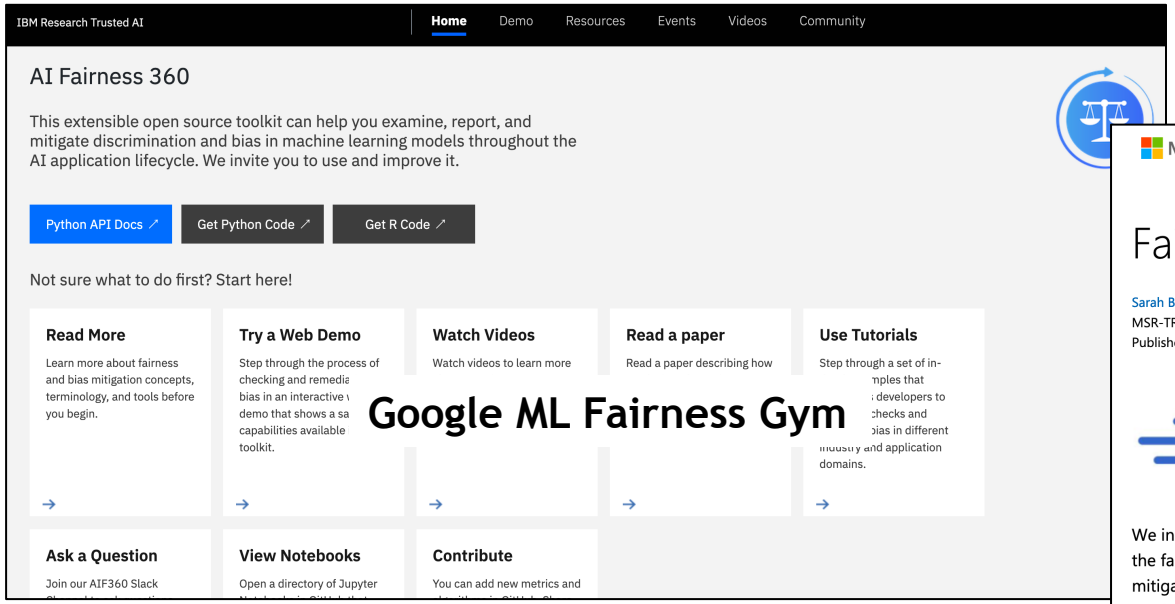


Who selects these criteria? Who defines what sensitive groups are?

Wattenberg, M. et al. (2021). Attacking discrimination with smarter machine learning, https://research.google.com/bigpicture/attacking-discrimination-in-ml/ (accessed Jan 07, 2025)

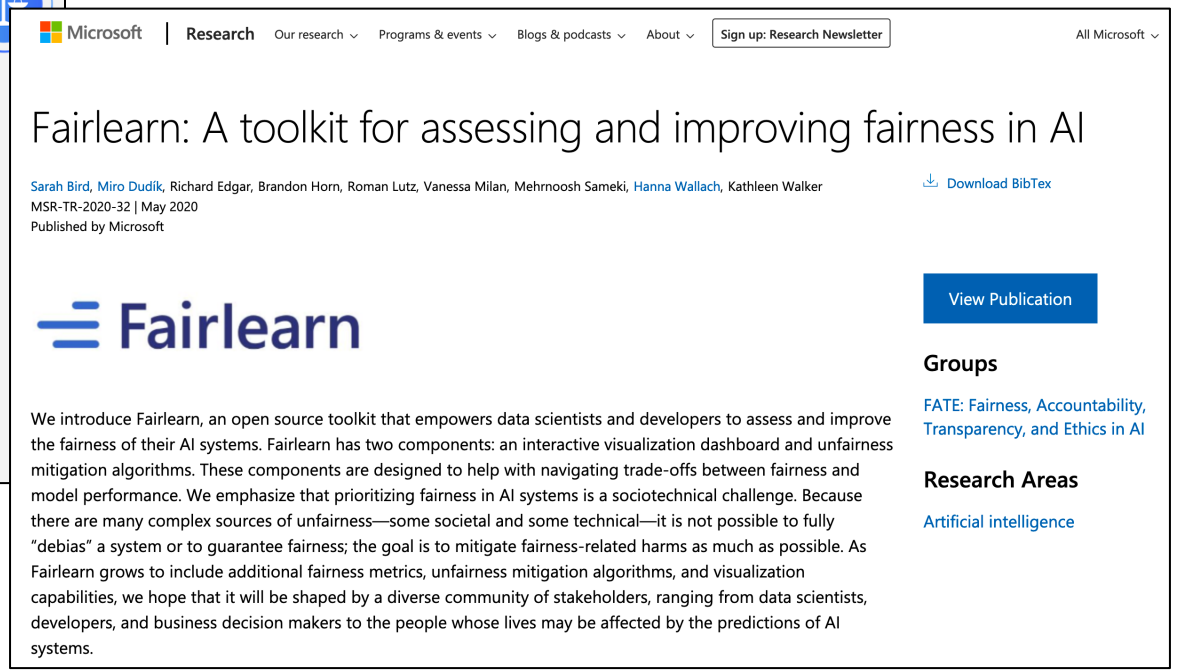
Tackling the interdisciplinary challenge of fairness
How to get started? Some helpful libraries

IBM AI Fairness 360 [1]



Google ML Fairness Gym

Microsoft Fairlearn [2]



IBM Research Trusted AI, AI Fairness 360 Toolkit, https://aif360.mybluemix.net (accessed Dec 07, 2024) [1]
Microsoft Research, Fairlearn: A toolkit for assessing and improving fairness in AI, https://www.microsoft.com/en-us/research/project/fairlearn (published May 2020) [2]

Tackling the interdisciplinary challenge of fairness
Let's ask stakeholders what they consider fair

Own Research

- Many formal fairness metrics have been developed—but what do **people actually consider *fair***?
- Fairness perceptions of AI decisions are **highly subjective** and **context-dependent**! [1]
  - ➢ How can we combine conflicting fairness perceptions?

> *"Life is unfair but remember sometimes it is unfair in your favor."*
>
> Peter Ustinov

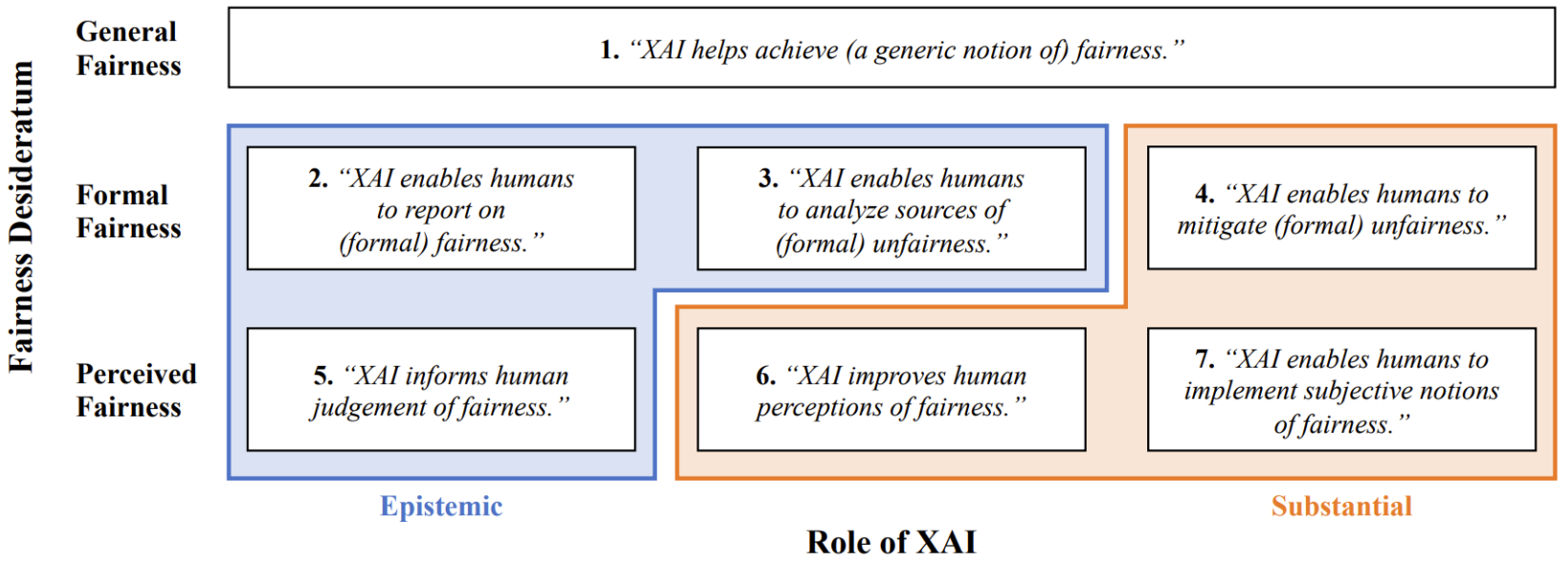
- **Demographic parity** seems to match respondent's ideas of fairness (n=100) [2]
- Stakeholders approve of **fairness through awareness** after explaining how sensitive features are used (n=58) [3]
- Decision-makers rely on AI recommendations that align to their personal fairness perceptions (n=280) [4]
  - ➢ Should we maximize perceived fairness?

⚠ **Fairness ideas vary** and affected people should be able to form **appropriate** fairness perceptions

Starke, C.; Baleis, J.; Keller, B.; Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. Big Data & Society, 9(2) [1]
Srivastava, A.; Heidari, H.; Krause, A. (2022). Mathematical Notions vs. Human Perception of Fairness. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining [2]
Nyarko, J.; Goel, S.; Sommers, R. (2022). Breaking Taboos in Fair Machine Learning: An Experimental Study. In: Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '21 [3]
Zipperling, D., Deck, L., Kolminsky-Rabs, V., Lanzl, J., Kühl, N. (2025). Bias Alignment in Human-AI Teams: Effects on Decision-Making. Work in progress. [4]

Tackling the interdisciplinary challenge of fairness
Will Explainable AI solve all fairness problems?

Own Research



Claims on the relationship between XAI and fairness are often overly **optimistic, vague, and simplistic.**

Many fairness desiderata related to XAI **lack normative grounding**, particularly regarding the role of sensitive features.

Many fairness desiderata are **poorly aligned with actual capabilities** of XAI disregarding sociotechnical and cognitive limitations.

XAI can be **one among many tools** to uncover unfairness but is not suited for **"proofs of fairness"**

Deck, L., Schoeffer, J., De-Arteaga, M., Kühl, N. (2023). A Critical Survey on Fairness Benefits of XAI. FAccT '224: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.

Tackling the interdisciplinary challenge of fairness
Will Explainable AI solve all fairness problems?

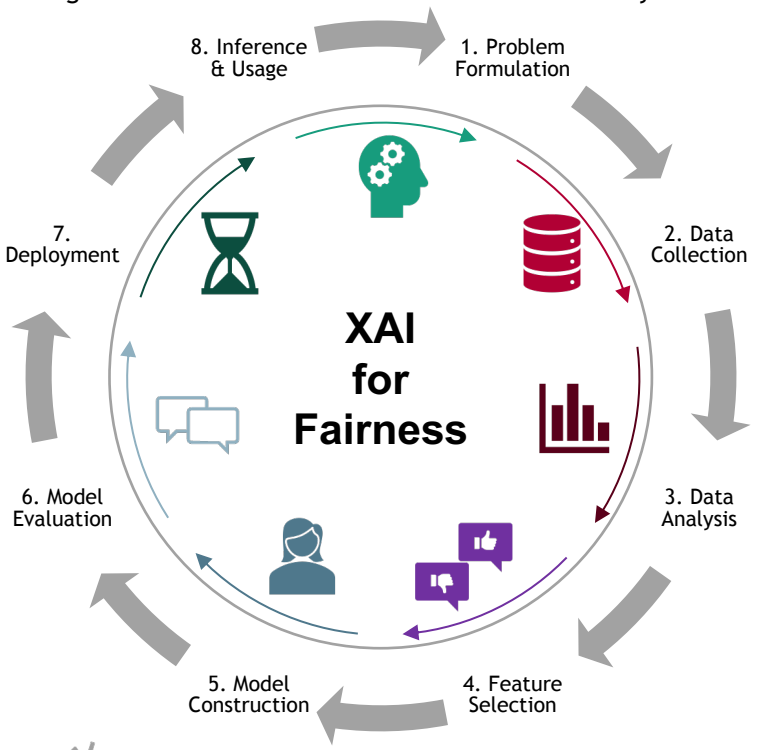
Own Research

Fairness Understanding
Enrich developers' understanding of fairness objectives with insights from all fairness endeavors across the lifecycle.

Long-Term Fairness
Monitor fairness to uncover fairness drifts and unfair downstream impacts of model deployment.

Empowering Fairness
Provide affected parties with actionable explanations to denounce unfair treatment and improve their outcome.

Fairness with Human Oversight
Provide tailored information to the human-in-the-loop for effective oversight or reliance of human decision-makers.



Data Fairness
Compute descriptive statistics to reveal undesirable data properties and guide strategies to counteract these.

Formal Fairness
Test and analyze formal fairness to reveal undesirable model properties and to guide mitigation strategies.

Perceived Fairness
Explain the model and the decision-making process to improve fairness perceptions and increase acceptance.

Informational Fairness
Provide truthful, understandable, and relevant information **about all fairness desiderata across the AI lifecycle.**

Deck, L., Schomäcker, A., Speith, T, Schöffer, J. Kästner, L. Kühl, N. (2024). Mapping the Potential of Explainable AI for Fairness Along the AI Lifecycle. European Workshop on Algorithmic Fairness (EWAF'24)..

1. What does the future look like?
2. The problem with accuracy
3. Can’t we just ignore sensitive information?
4. Tackling the interdisciplinary challenge of fairness
5. **Outlook: We still have a long way to go**

Outlook: We still have a long way to go
Fairness, accuracy, and transparency pose a complex tradeoff

Accuracy
- Simple measure
- Reflects economic value
- Reinforces social inequality

Formal Fairness
- Might sacrifice "historical" accuracy [1]
- Context-sensitive, conflicting measures

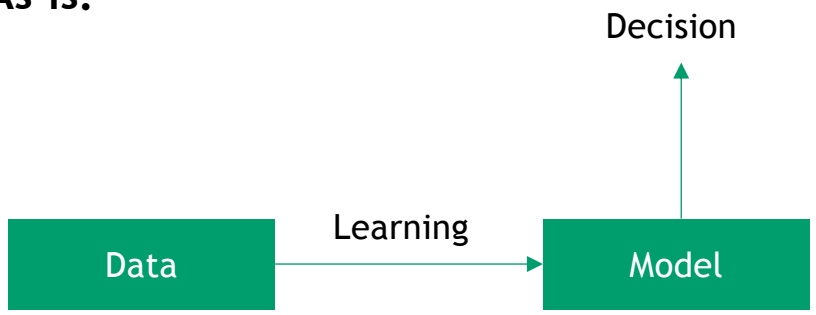
Transparency
- Allows insights into "right kind" of accuracy and sensitive attributes
  - Might sacrifice accuracy and formal fairness [2]
- Might provide informational fairness

Rodolfa, T. et al. (2021). Machine learning for public policy: Do we need to sacrifice accuracy to make models fair? [1]
Bell, A. et al. (2022): It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 248-266 [2]

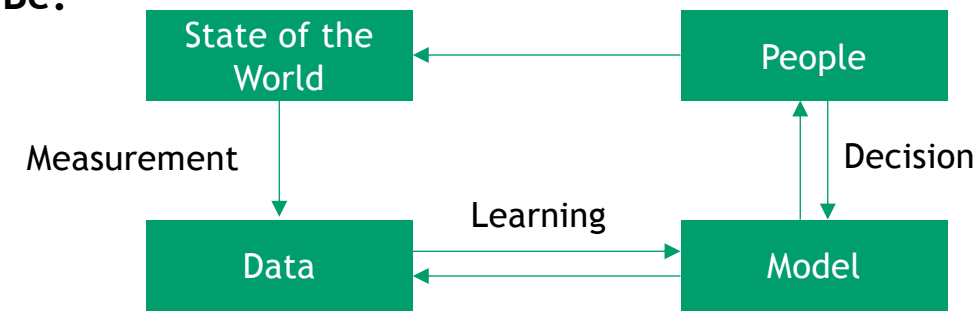
Outlook: We still have a long way to go
A broader perspective on Ethics of AI

- Fairness in AI-based technology is a **holistic societal** problem that AI research alone **cannot** solve
- Instead, we need to also consider feedback loops changing the state of the world (“**socio-technical system**”)

As-Is:



To-Be:



- **Informational & Interpersonal Fairness**: How should affected people be treated when receiving an AI-driven decision?
  - Right to explanation? Right to contest? Human point of contact? Guidance on recourse?
- **Selective Labels**: data on university success only includes students who have been accepted in the first place
  - Should we include underrepresented minorities to improve imbalanced data?
- **Prediction vs. Intervention**: AI is not the only tool to affect the state of the world, we can also act directly
  - Should we predict criminality to lock up suspects OR support people to not even become criminals in the first place?

Hardt, M. (MLSS 2020), https://www.youtube.com/watch?v=Igq_S_7IfOU (accessed Jan 07, 2025).

Summary

- AI can **repeat and even reinforce** existing patterns of unfairness and social inequality.
- Try to be **explicit** about your **objective,** your **moral considerations,** and the **impact on real people's lives**.
- Omitting sensitive features is **not** solving the problem → implement sensitive features **responsibly and meaningfully**.
- There are **many notions of fairness**, but you will have a hard time finding a "**fairness certificate**".
- **Explainable AI** is a powerful tool, but its strengths and weaknesses depend on your **objective** and your **stakeholders**.
- Take your time to consider the **consequences** and the **socio-technical system** in which your model will be deployed**.**

Thank you for your participation! Don’t hesitate to contact us!

Prof. Dr. Niklas Kühl

kuehl@uni-bayreuth.de

Luca Deck

luca.deck@uni-bayreuth.de