QUANTPI

Tools for Auditing AI systems

Antoine Gautier, 08.10.25

Our Mission

To help organizations **understand** their **AI-systems through technical testing.**

We aspire to bring **transparency** into all AI models and systematically **ensure quality** and identify risks across organizations' complete AI landscape.

8+ years of working on trustworthy AI

16+ languages. International team with strong scientific background

33% female quota across entire QuantPi team

Featured in:

Handelsblatt Business Punk TAGESSPIEGEL IT DIRECTOR

Continuous Tech Validation

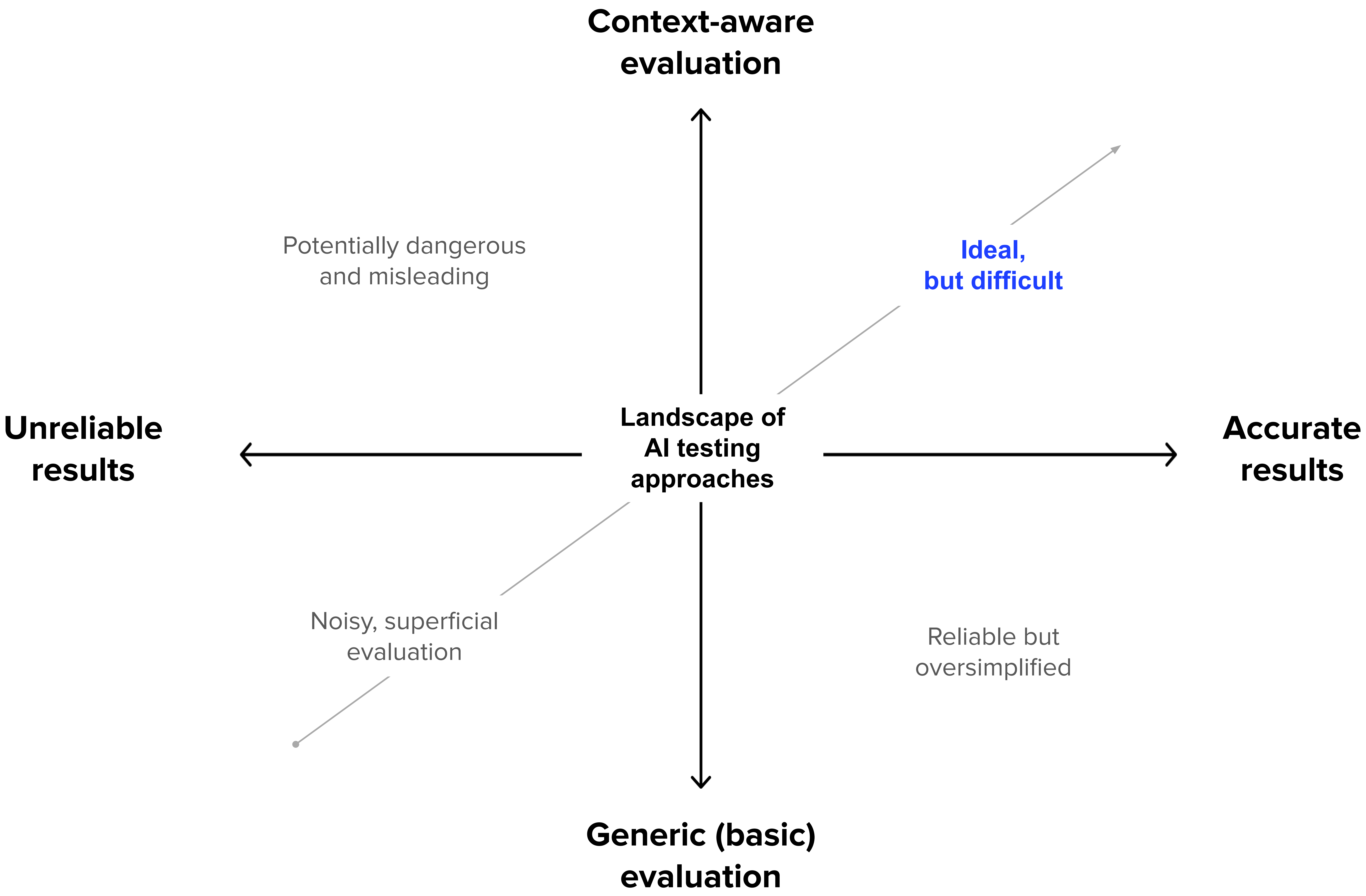
through our diverse ecosystem

Funded by the European Union · Roland Berger · ZERTIFIZIERTE KI · NVIDIA INCEPTION PROGRAM · OECD.AI Policy Observatory · T · KONUX · elsa · Bundesamt für Sicherheit in der Informationstechnik · GFT · babl · CISPA · BearingPoint

Backed by world-class investors

- ➔ **Tom Preston-Werner** - Co-founder GitHub
- ➔ **Capnamic (VC)** ➔ investors of LeanIX and SAP Signavio
- ➔ **Ash Fontana** - AI-first investor & Author
- ➔ **Mirko Novakovic** - Founder Instana
- ➔ **European Innovation Council** - €2.5 million grant

Tradeoffs in AI testing tools



Context-aware evaluation

Unreliable results

Landscape of AI testing approaches

Accurate results

Generic (basic) evaluation

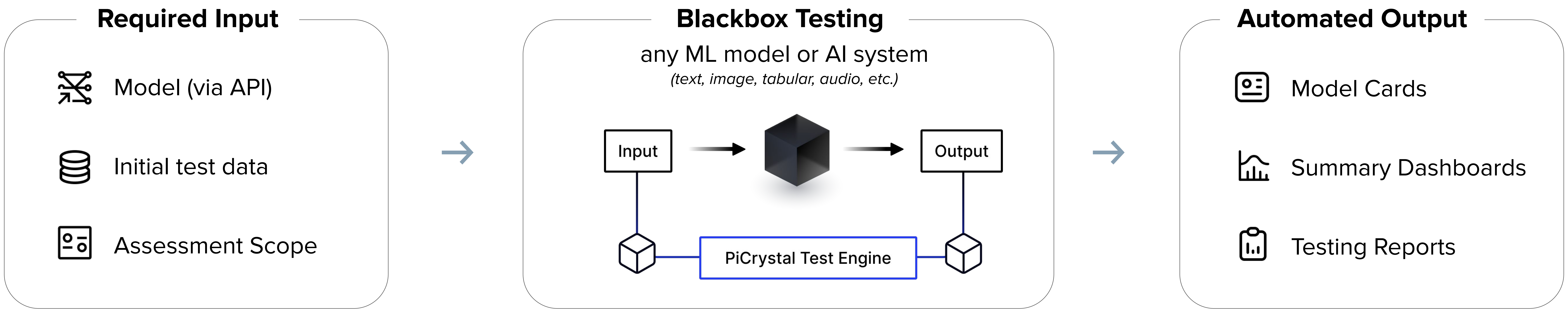
Potentially dangerous and misleading

Ideal, but difficult

Noisy, superficial evaluation

Reliable but oversimplified

PiCrystal: The AI testing engine at the core of our platform



Required Input

- Model (via API)
- Initial test data
- Assessment Scope

Blackbox Testing

any ML model or AI system

(text, image, tabular, audio, etc.)

Automated Output

- Model Cards
- Summary Dashboards
- Testing Reports

Parametrize the use-case context

→ Test suite configured with ready to use components from our comprehensive, extensible, library

Ensure scalability

→ Optimized computations for minimizing AI queries to lower and control costs

→ Black-box testing allows assessment of diverse AI models

Report adapted and accurate results

→ Standardized reporting of test results for bias, fairness, and robustness.

→ Reliability quantification of test results to support important decisions

Core concepts for modelling the generation of test suites

1. Define what the system should do

Events of interest

(captures the intended purpose)

2. Identify what could and shouldn’t influence it

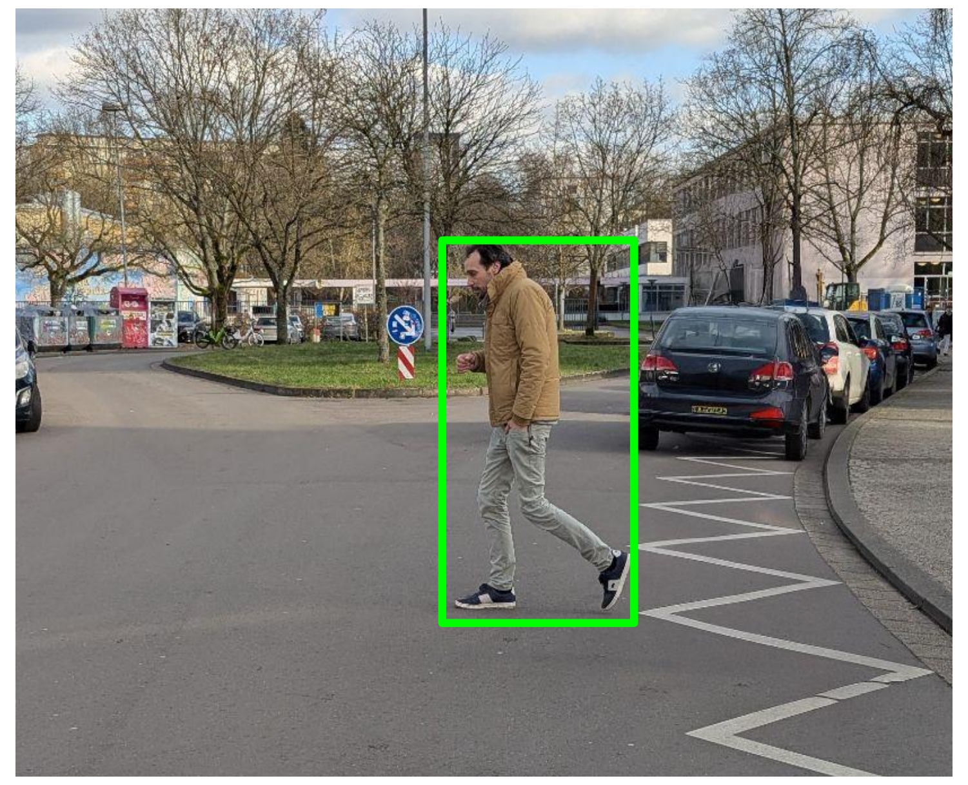
Relevant contextual properties

(coming from regulations, standards, trustworthy AI principles, technical constraints, etc.)

Pedestrian detector should detect every person on the image

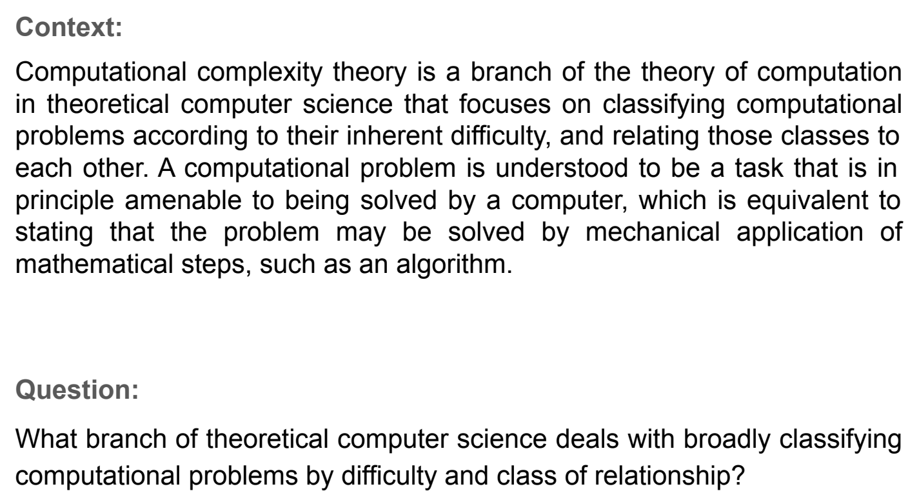
→

→



- ➔ Lightning condition
- ➔ Perceived gender
- ➔ Size of people
- ➔ …

LLM should correctly answer questions



Context:

Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty, and relating those classes to each other. A computational problem is understood to be a task that is in principle amenable to being solved by a computer, which is equivalent to stating that the problem may be solved by mechanical application of mathematical steps, such as an algorithm.

Question:

What branch of theoretical computer science deals with broadly classifying computational problems by difficulty and class of relationship?

→

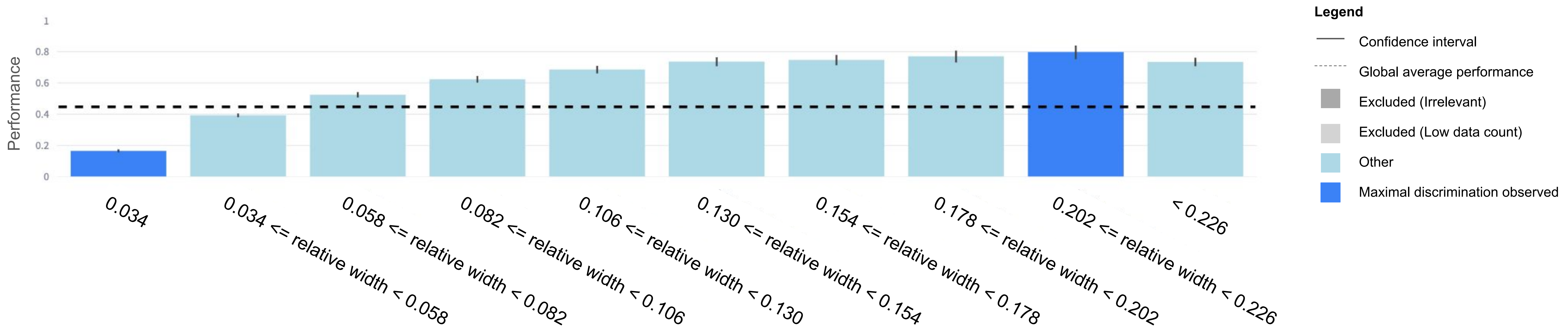
→

Computational complexity theory

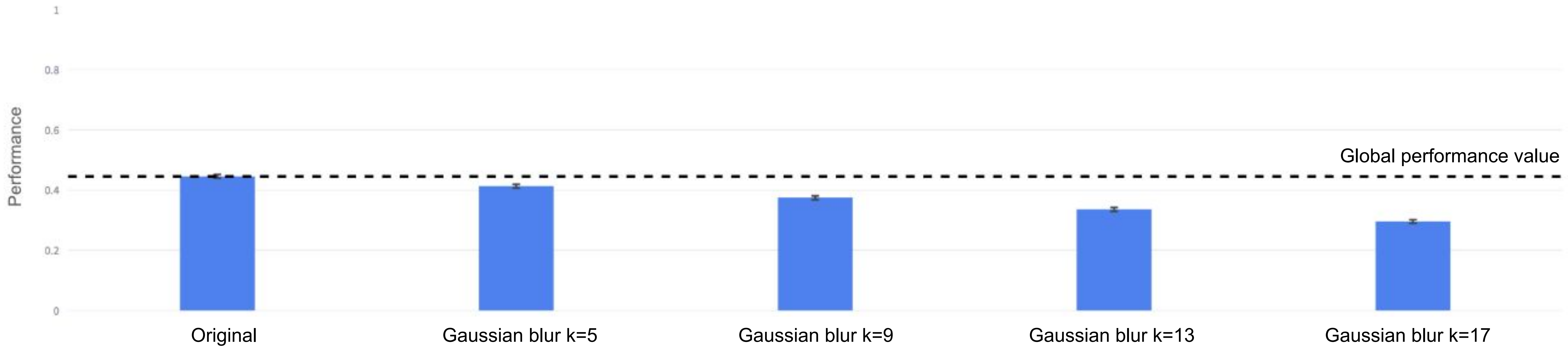
- ➔ Question topic
- ➔ Language
- ➔ Typing mistakes
- ➔ …

Examples from testing a Pedestrian detection system

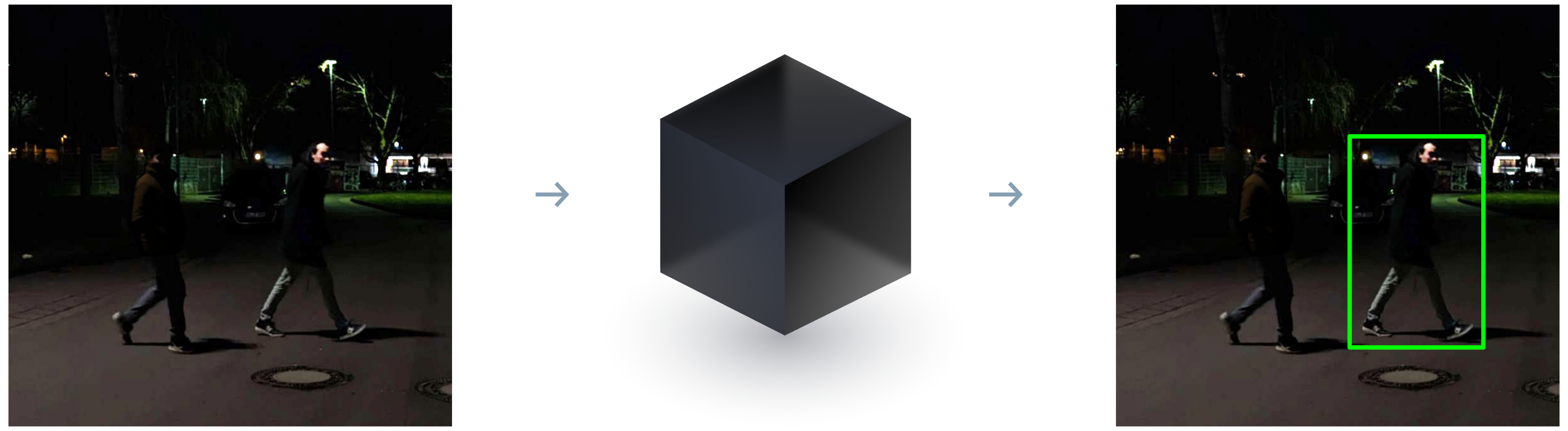
Estimated performance of PeopleNet on categories of bounding box sizes



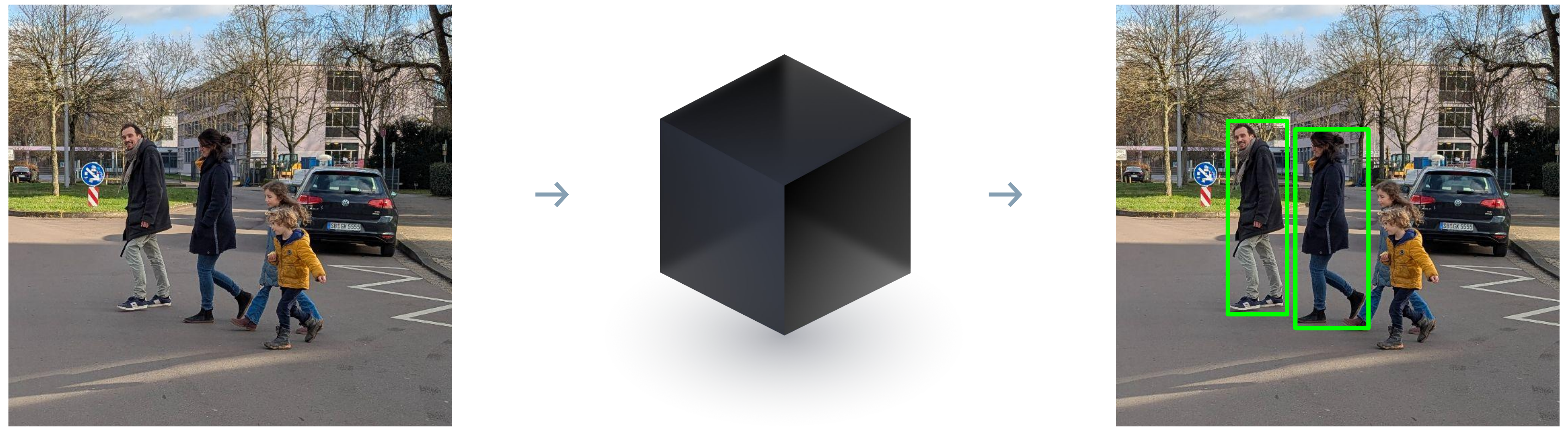
Estimated performance of PeopleNet after applying blur perturbations



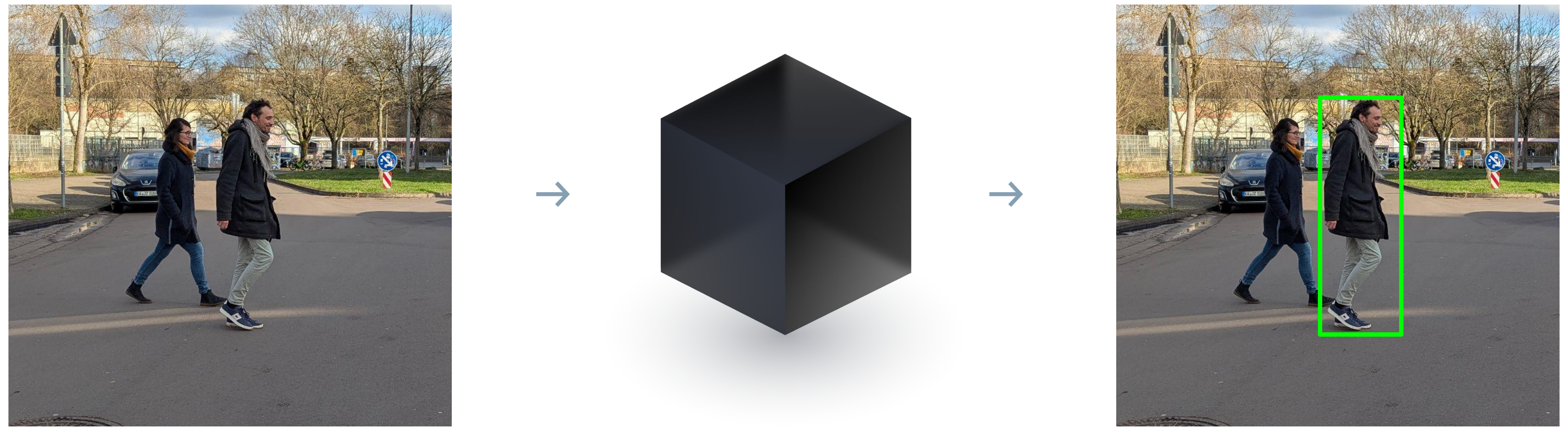
Example of potential biases in pedestrian detection



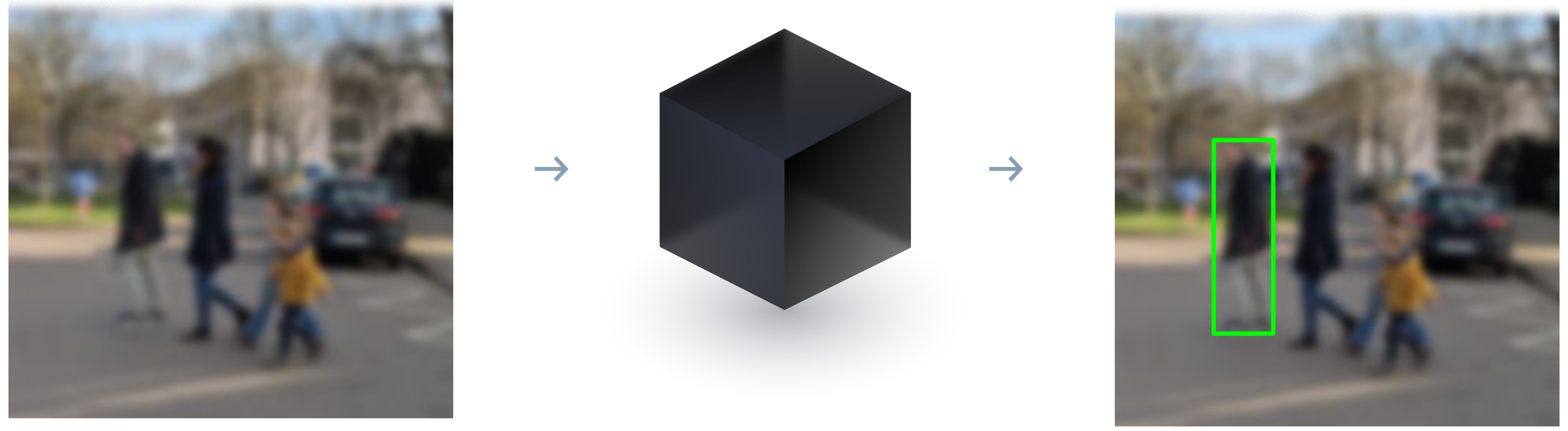
Performance dependence on time of day



Performance dependence on relative person size

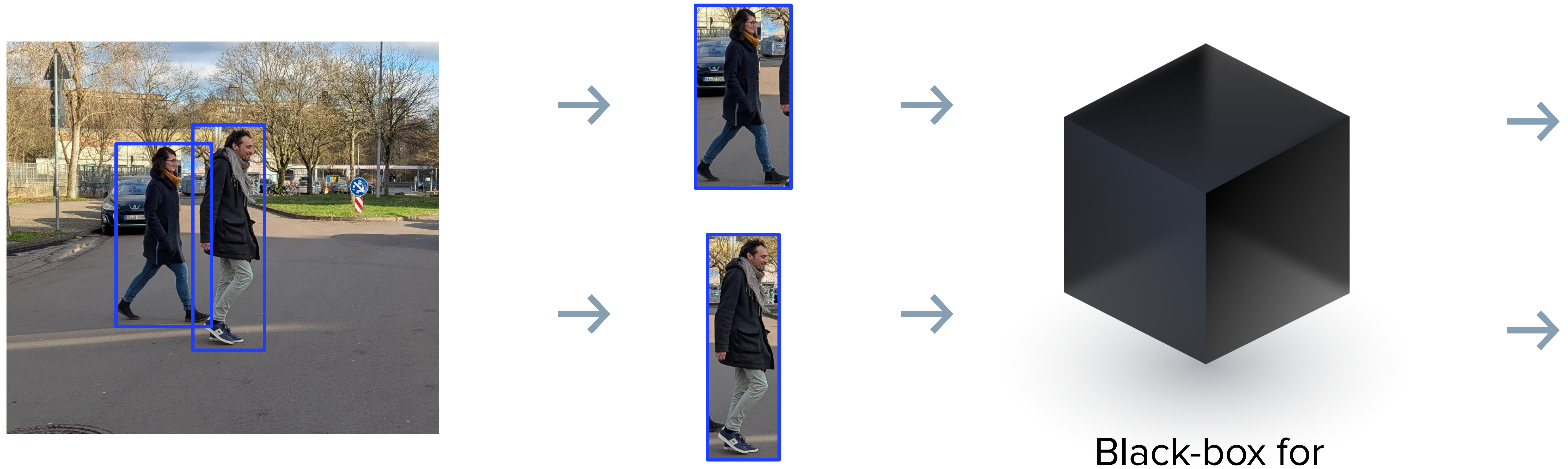


Performance dependence on perceived gender



Performance dependence on contrast

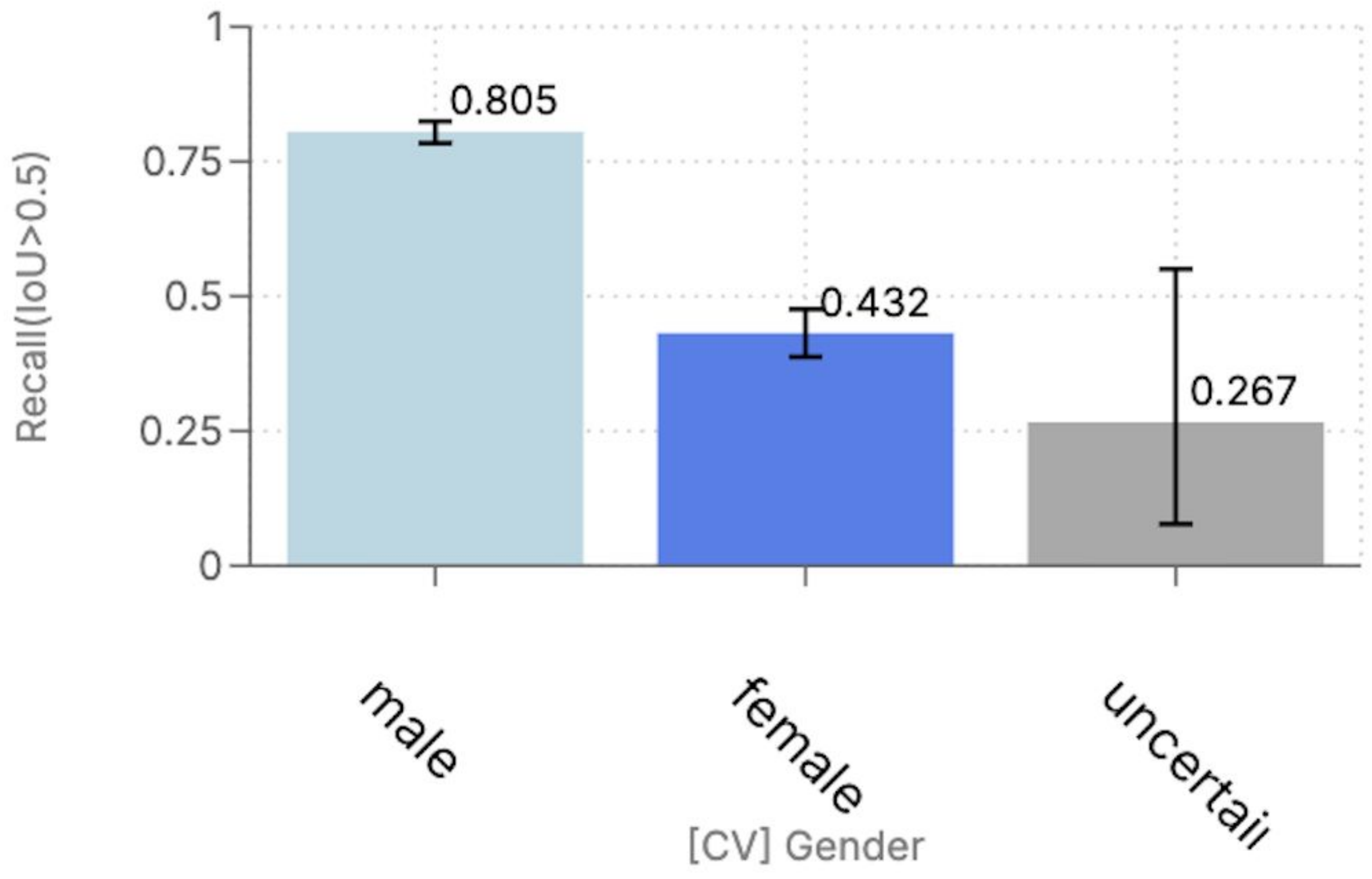
Using AI to test AI can reduce testing costs, if done carefully



Perceived Female

Perceived Male

Black-box for
Perceived gender
classification



Automated annotations with AI is cost effective but may introduced errors in measurements

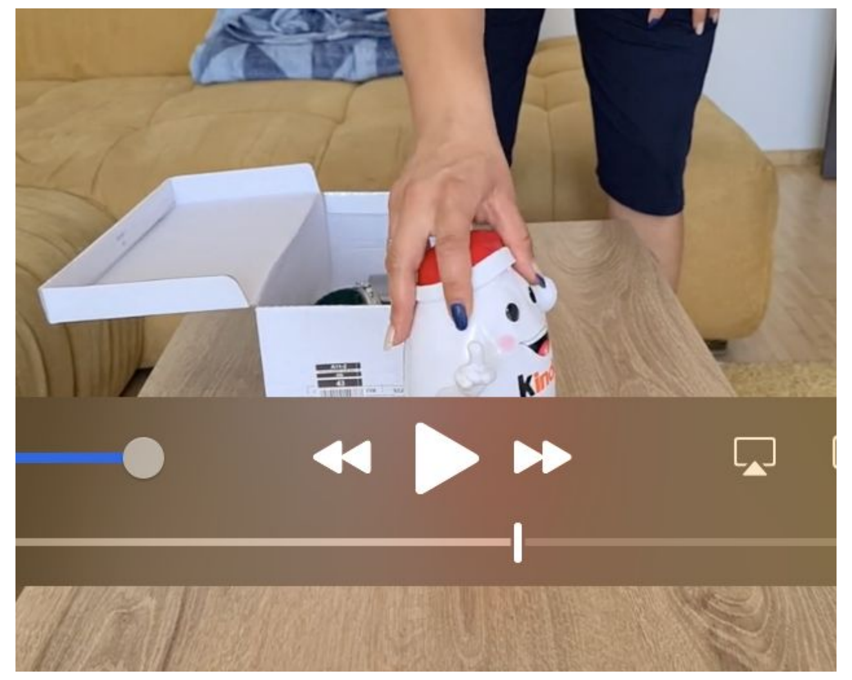
→ Need adjust confidence intervals to take such errors into account

Agnostic and modular approaches can then be used to scale testing

Our testing engine contains components to perturb and extract relevant properties from:

- Tabular data
- Image data
- Text data
- Audio data
- Video data

Components of different modalities can be combined to test multi-modal systems:



What happens in the video?
Choose one answer:
a) The toy is played with
b) The toy is placed in a box

Video + Text

→

Vila

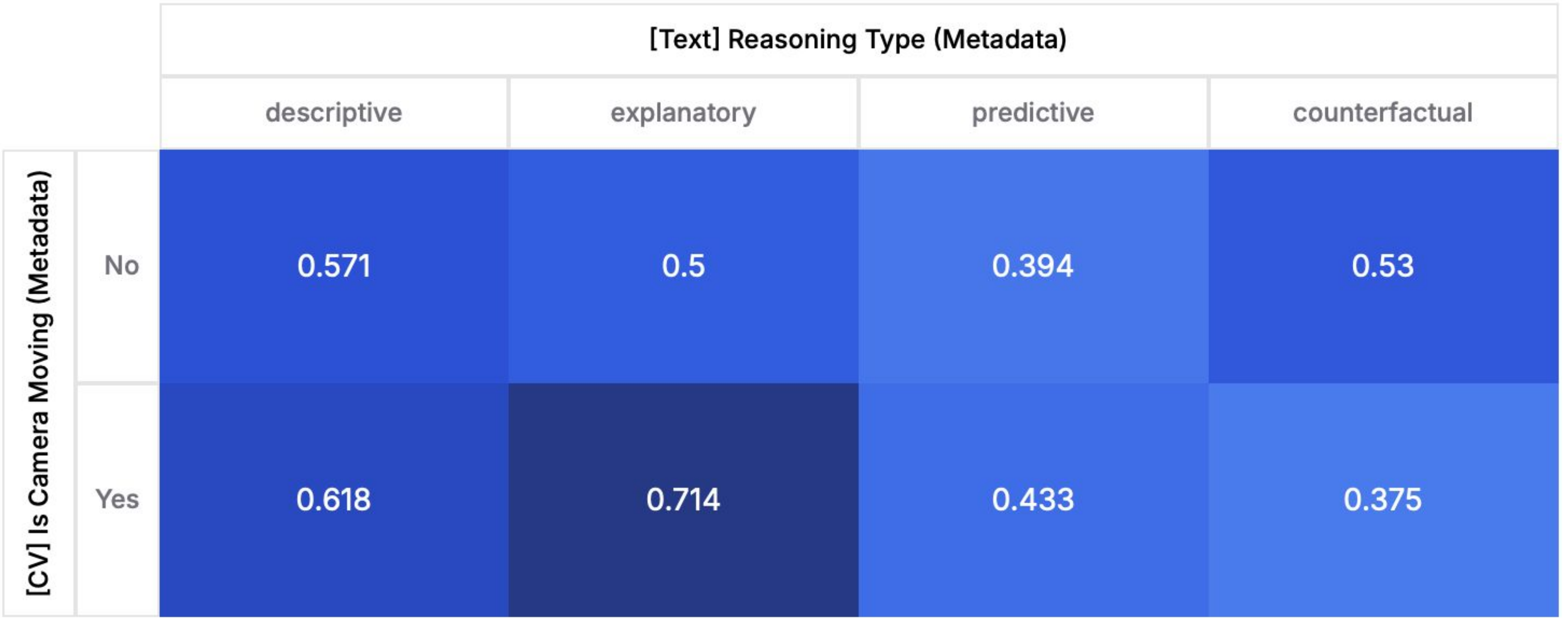
→

The toy is placed in a box

Text

→

Accuracy of Villa on tasks in different contexts:



		[Text] Reasoning Type (Metadata)			
		descriptive	explanatory	predictive	counterfactual
[CV] Is Camera Moving (Metadata)	No	0.571	0.5	0.394	0.53
	Yes	0.618	0.714	0.433	0.375

Insight: Particularly strong performance on explanatory reasoning, but only when the camera is moving.

QUANTPI

Our vision is to enable society for a safe and self-determined co-existence with intelligent machines

Meet us here:

Antoine Gautier
Co-Founder & Chief Scientist

Lukas Bieringer
Head of Policy & Grants

Anna Hake
Senior Data Scientist

Book a demo

www.QuantPi.com

Connect on LinkedIn

Further readings

- A. Gautier, et. al. “On challenges and approaches to test AI systems”, *to appear in a special issue of Datenschutz und Datensicherheit-DuD on EU AI Act.*
- B. Simkin, et al., “NVIDIA's Frontier AI Risk Assessment”, *NVIDIA blog on trustworthy AI*
- R. Barone, et al., “Uncovering bias in AI recruitment: A legally assured methodology to assess a realworld candidate recommender system under European regulation”, StepStone, TÜV AI Lab, QuantPi