



The Standards People

Trusted AI for Business

Presented by: **Alessandro Curioni,**
IBM Fellow, Vice President
Europe and Director,
IBM Research Zurich

For: **ETSI Summit on Artificial**
Intelligence

AI is...

Artificial intelligence



Machine learning

Neural networks

Deep learning

Specialized AI works incredibly well

Language
translation

Speech
transcription

Language
processing

Object
detection

Face
recognition

94%

of companies believe that AI is the key to competitive advantage

80%

of data is either inaccessible, untrusted or unanalyzed

81%

do not yet understand the data required for AI

1 in 20

companies have extensively incorporated AI into offerings and processes

60%

See compliance as a barrier due to a lack of trust in AI outcomes

65%

do not fully trust their own organizations analytics

The evolution of AI

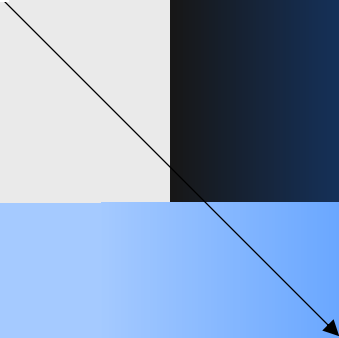
General AI
Revolutionary

We are here

Broad AI
Disruptive and pervasive

2050 and beyond

Narrow AI
Emerging



The evolution of AI

Narrow AI

Deep learning

Single-task, single-domain,
with superhuman accuracy

Requires large amounts
of labeled data

Broad AI

Learning + reasoning

Multi-task, multi-domain,
multi-modal

Learns with
much less data

General AI

True neuro-AI

Cross-domain learning
and reasoning

Broad
autonomy

AI for Enterprise



Today's Enterprise AI



Core AI

Imbued with the characteristics of human trust

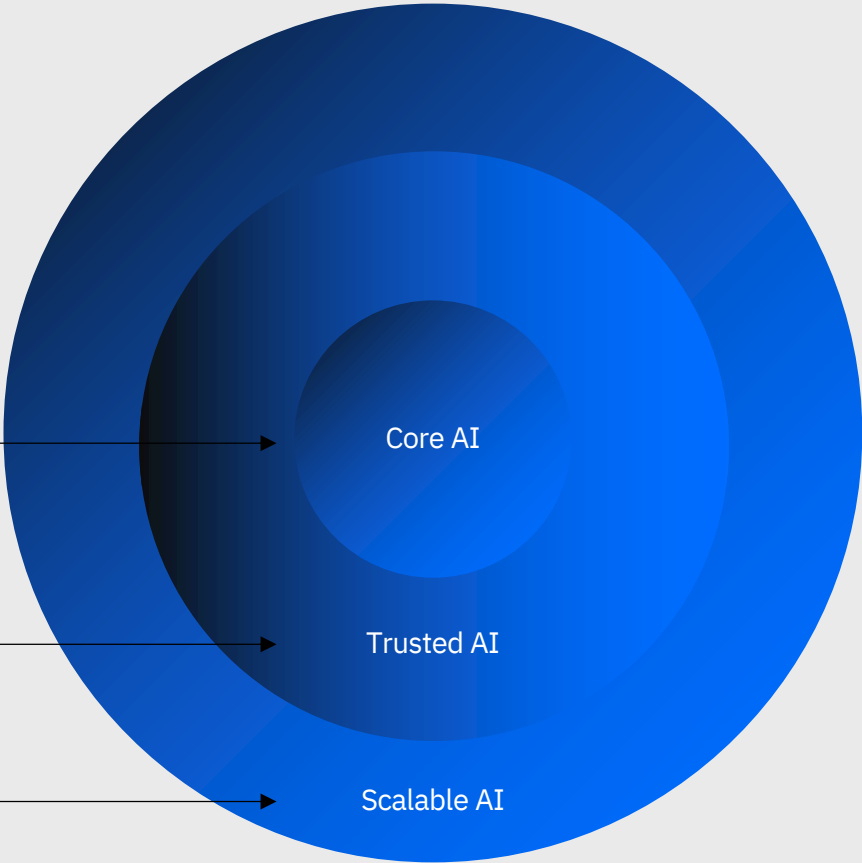


Trusted AI

Manages and operates Trusted AI and its lifecycle



Scalable AI



AI for Enterprise



Today's Enterprise AI



Core AI

Imbued with the characteristics of human trust

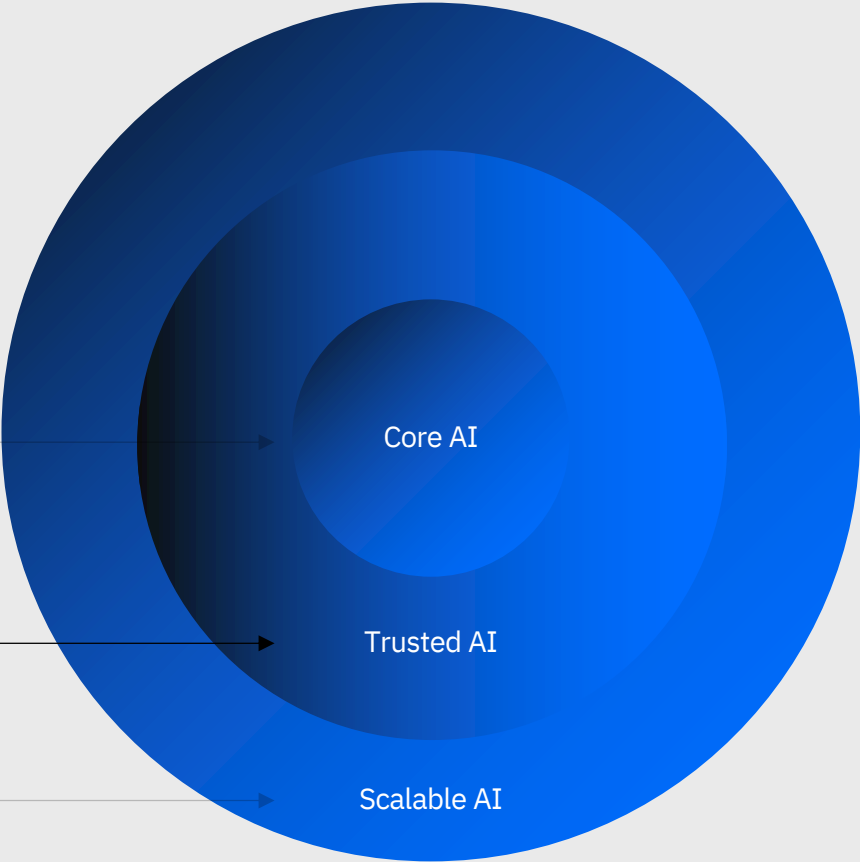


Trusted AI

Manages and operates Trusted AI and its lifecycle



Scalable AI



What does it take
to trust a decision
made by a machine?

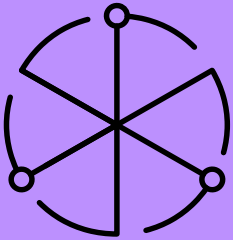
The pillars of trust

Is it fair?

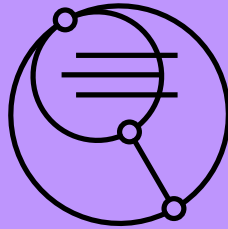
Is it easy to understand?

Is it secure?

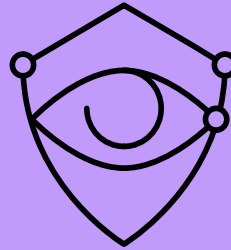
Is it accountable?



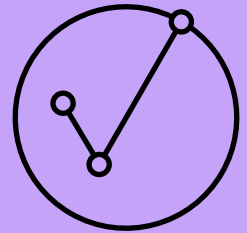
Fairness



Explainability

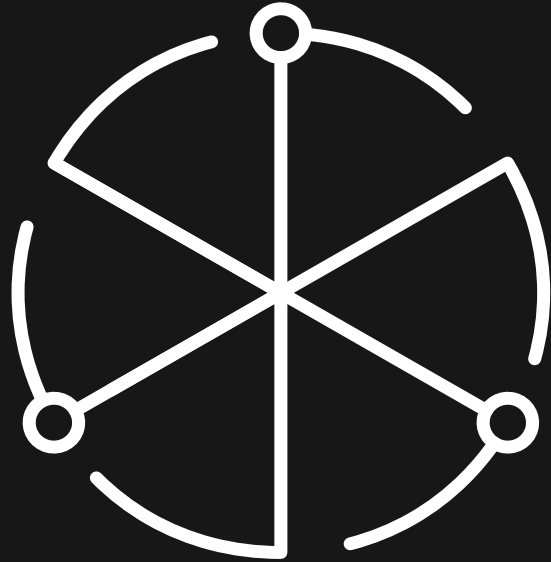


Adversarial Robustness



Transparency

Making AI Fair



AI Fairness 360:

- ▶ 30+ fairness metrics/checkers
- ▶ 10 bias “mitigators”
- ▶ industry tutorials

The screenshot shows the IBM Research Trusted AI website for the AI Fairness 360 Open Source Toolkit. The navigation bar includes links for Home, Demo, Resources, Events, and Community. The main content area features a title, a descriptive paragraph, two buttons for API Docs and Get Code, a call to action for starting here, and a grid of six cards: Read More, Try a Web Demo, Watch a Video, Read a paper, Use Tutorials, and Ask a Question.

IBM Research Trusted AI | [Home](#) | [Demo](#) | [Resources](#) | [Events](#) | [Community](#)

AI Fairness 360 Open Source Toolkit

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

[API Docs ↗](#) [Get Code ↗](#)

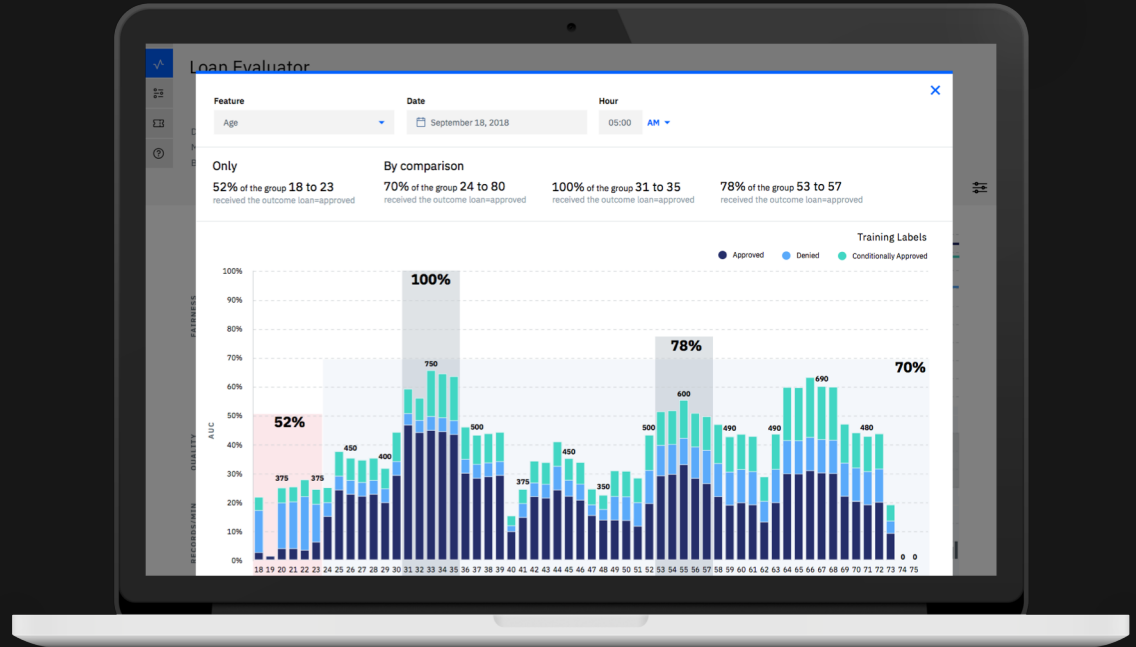
Not sure what to do first? Start here!

- Read More**
Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.
[→](#)
- Try a Web Demo**
Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.
[→](#)
- Watch a Video**
Watch a video to learn more about AI Fairness 360.
[→](#)
- Read a paper**
Read a paper describing how we designed AI Fairness 360.
[→](#)
- Use Tutorials**
Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.
[→](#)
- Ask a Question**
Join our AIF360 Slack Channel to ask questions, make comments and tell stories about how you use the toolkit.
[→](#)

Trusted AI

Fairness in action: Watson OpenScale

Operationalizes
and instruments
fairness into
enterprise-grade
workloads

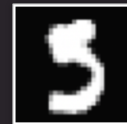


Making AI Explainable



Trusted AI

Algorithm: Contrastive explanations



We deduce that the image is “3” because,



we observe that curves are present,



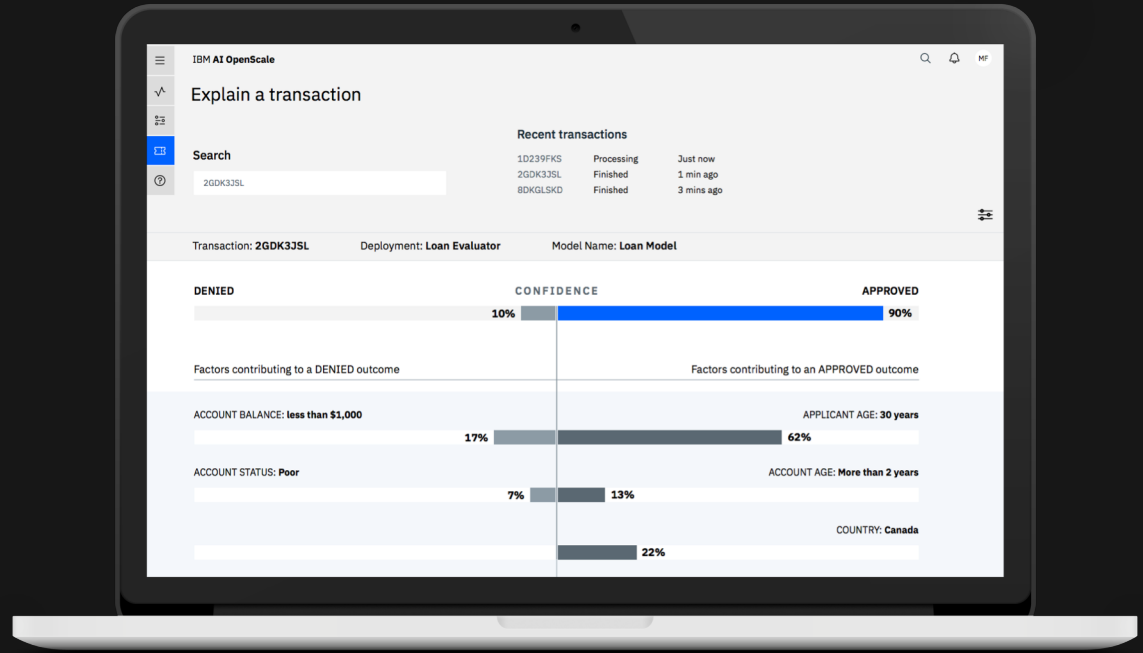
but a dash is absent.

Dhurandhar et al, “**Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives,**” NeurIPS 2018.

Trusted AI

Explainability in action: Watson OpenScale

Instruments
explanations into
enterprise-grade
workloads



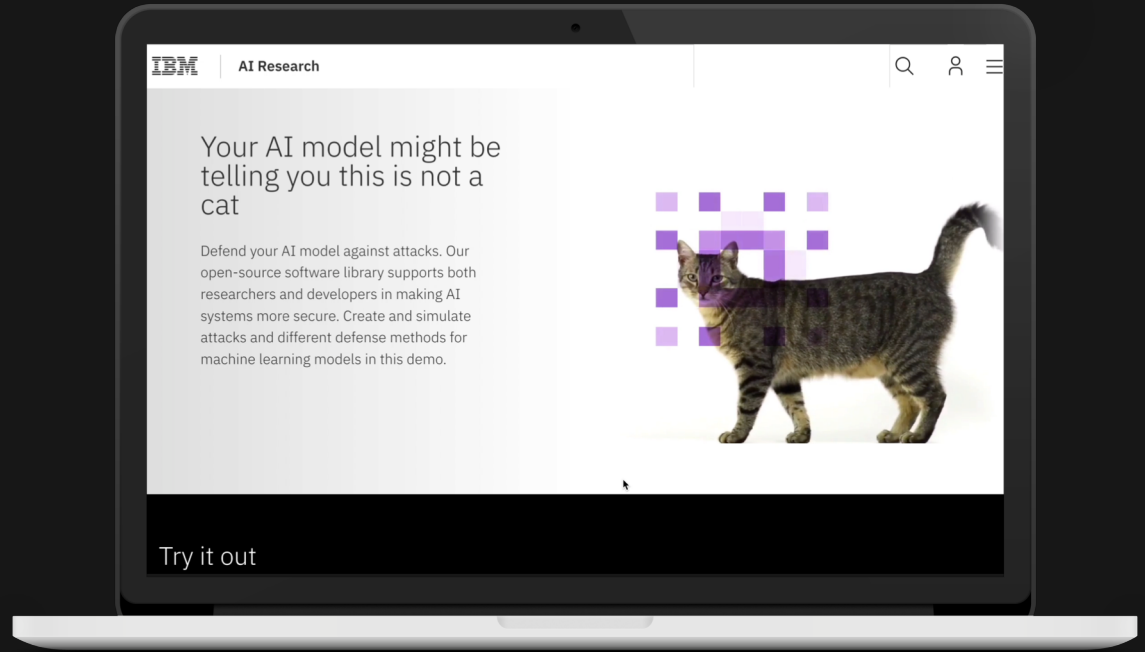
Making AI Secure



Trusted AI

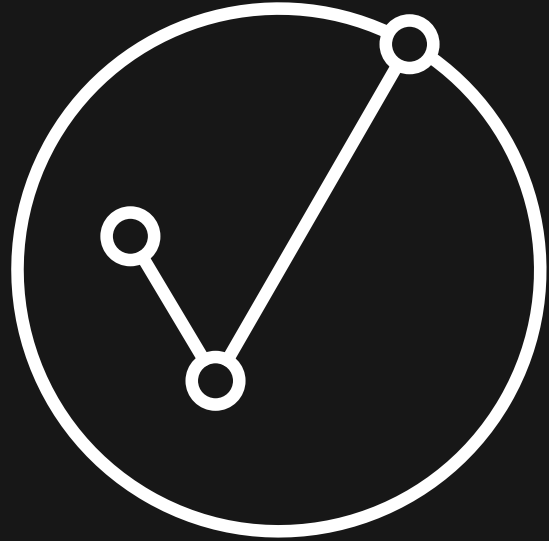
Adversarial Robustness Toolkit

The most
comprehensive
toolkit for
“attacking” and
defending AI



<https://art-demo.mybluemix.net/>

Making AI Transparent



Trusted AI

Factsheets for AI



The image shows a hand pointing to a Nutrition Facts label on a cereal box. The label is partially obscured by the hand and the box's texture. The text on the label is as follows:

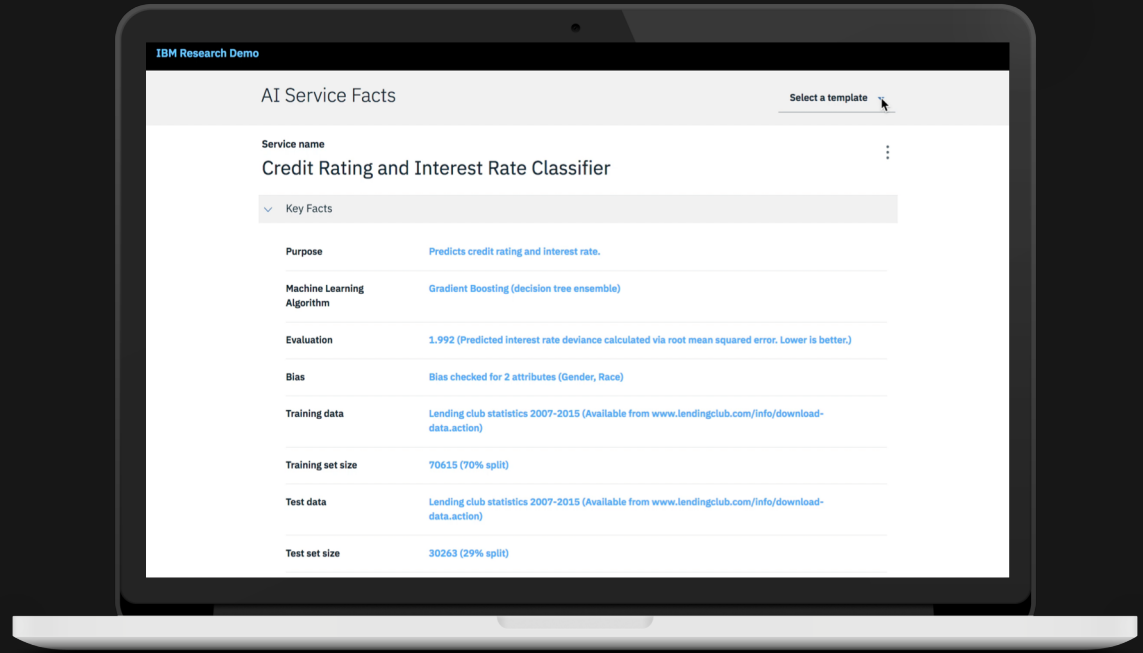
Nutrition Facts	
Per 3/4 cup (29 g)	
Amount	
Calories	110
	170
	% Daily Value
Fat 1 g*	6 %
Saturated 0.3 g	
+ Trans 0 g	
Cholesterol 0 mg	
Sodium 180 mg	
Carbohydrate 23 g	
Fibre 2 g	
Sugars 10 g	
Vitamin A	0 %
Vitamin C	0 %
	10 %
	30 %

Arnold et al, "FactSheets: Increasing Trust in AI Services through Supplier's Declaration of Conformity," <https://arxiv.org/pdf/1808.07261.pdf>

Trusted AI

Transparency in action: Factsheets for AI

A transparent
reporting
mechanism
for how an AI
system operates
and performs

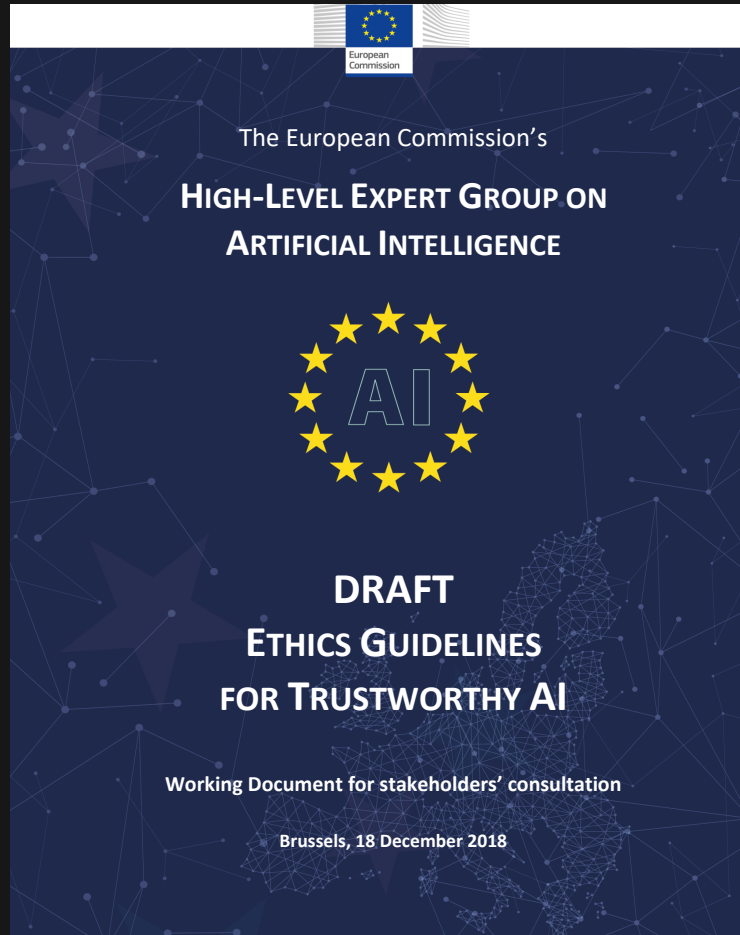


Trusted AI

Coming Soon

Ethics guidelines
for trustworthy AI

AI Ethics
Guidelines
presented by the
European
Commission's
High-Level Expert
Group on Artificial
Intelligence (AI
HLEG).



Building more
trustworthy AI systems
is not only a research
question, it's a
business imperative.

IBM®