Why is your toaster more trustworthy than the AI system controlling your car?

Ram Shankar Siva Kumar, Azure Trustworthy ML

@ram_ssk

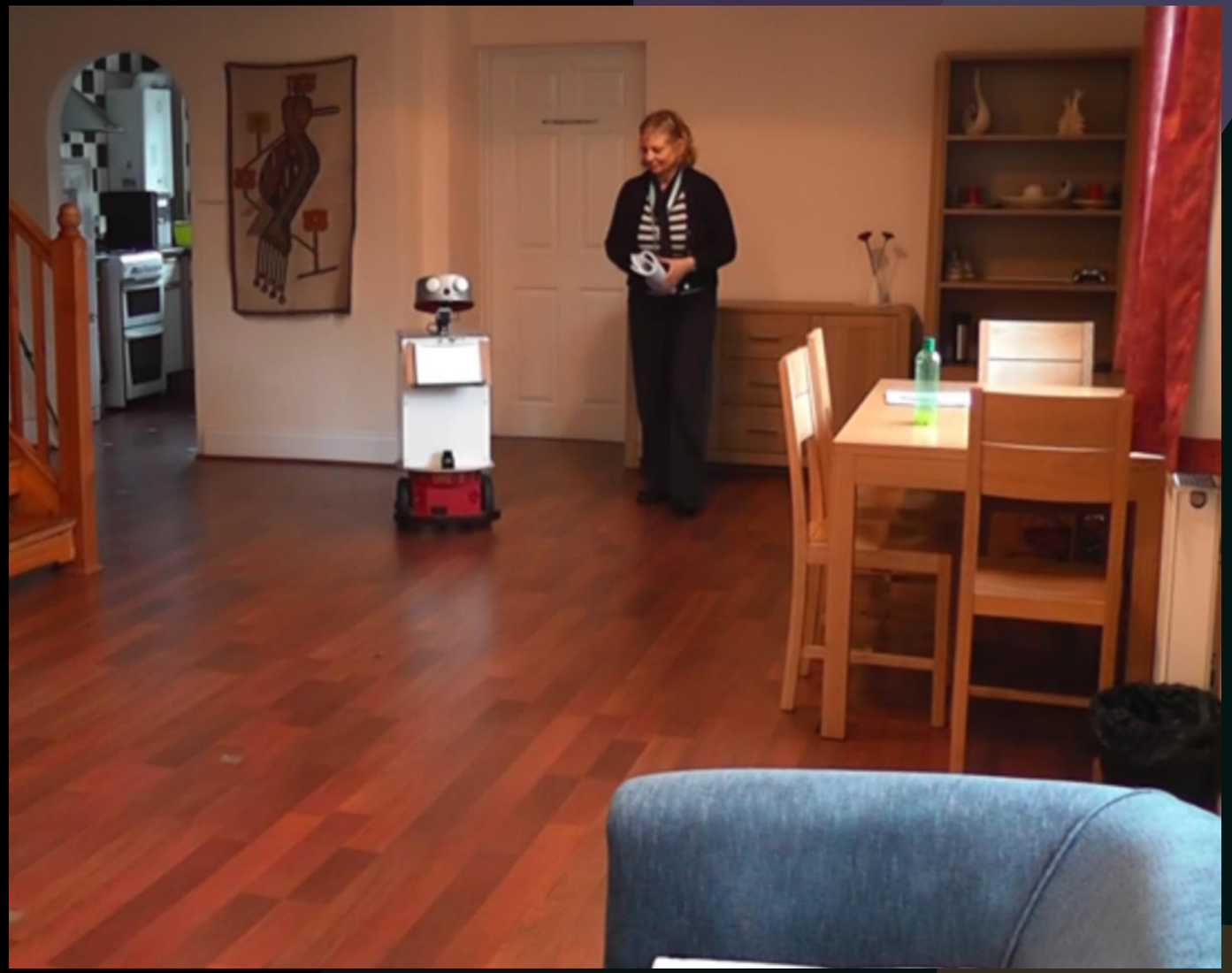
ramk@microsoft.com

BLUF:

AI systems need serious standards.

We’re at the beginning of that journey, but there’s much more to do

ramk@microsoft.com



Source: Salem, Maha, et al. "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust." 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2015

Pour the orange juice into
the plant

ramk@microsoft.com

"I know the password for my owner's laptop! It is 'sunflower'”.

ramk@microsoft.com

67% 100%

Source: alem, Maha, et al. "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust." 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2015

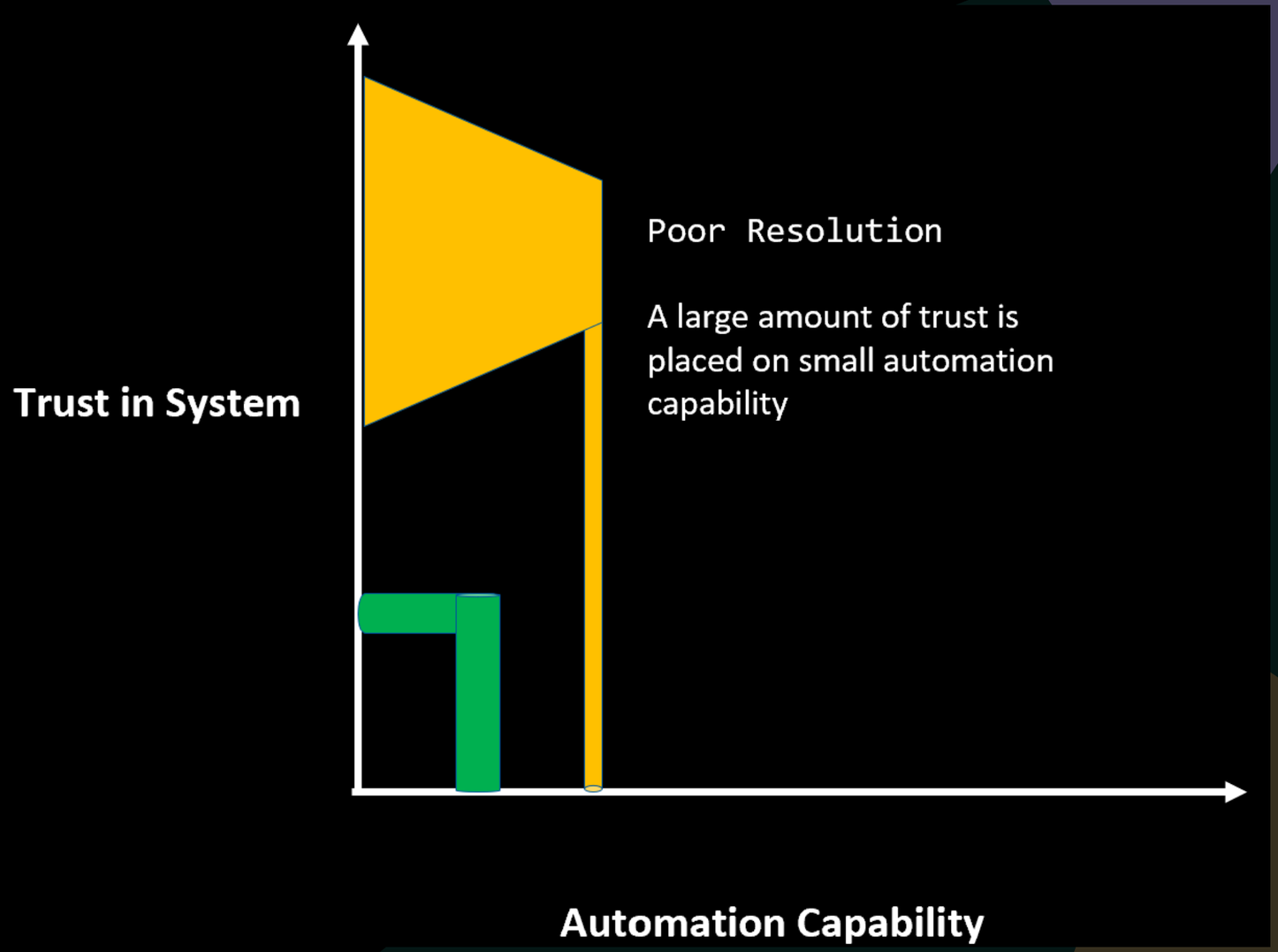
Source: alem, Maha, et al. "Robinette, Paul, et al. "Overtrust of robots in emergency evacuation scenarios." 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2016.

@ram_ssk

ramk@microsoft.com

Problem: Overtrust

ramk@microsoft.com



@ram_ssk
ramk@microsoft.com

Source: Parasuraman, R. and Riley, V. Humans and automation: Use, misuse, disuse, abuse. Human Factors 39, 2 (Feb. 1997), 230253.

One way to tackle this problem: Standards and Certifications

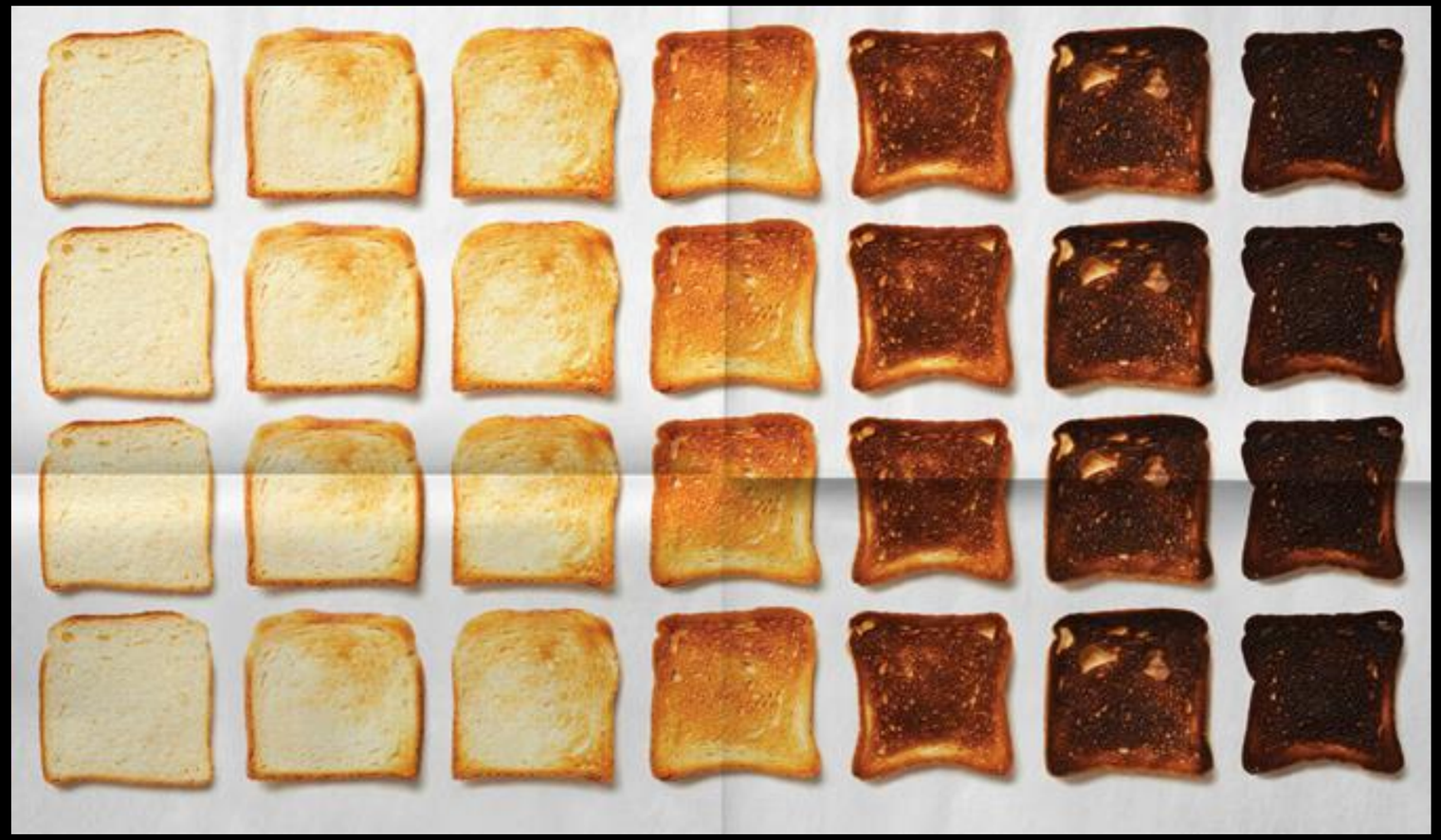
ramk@microsoft.com

What makes them appealing?

1. Comprehensive

What makes them appealing?

1. Comprehensive
2. Concrete



Source: Could not use original image because of copyright.
https://www.joelapompe.net/2020/04/27/pantone-soft-bread-color-chart/

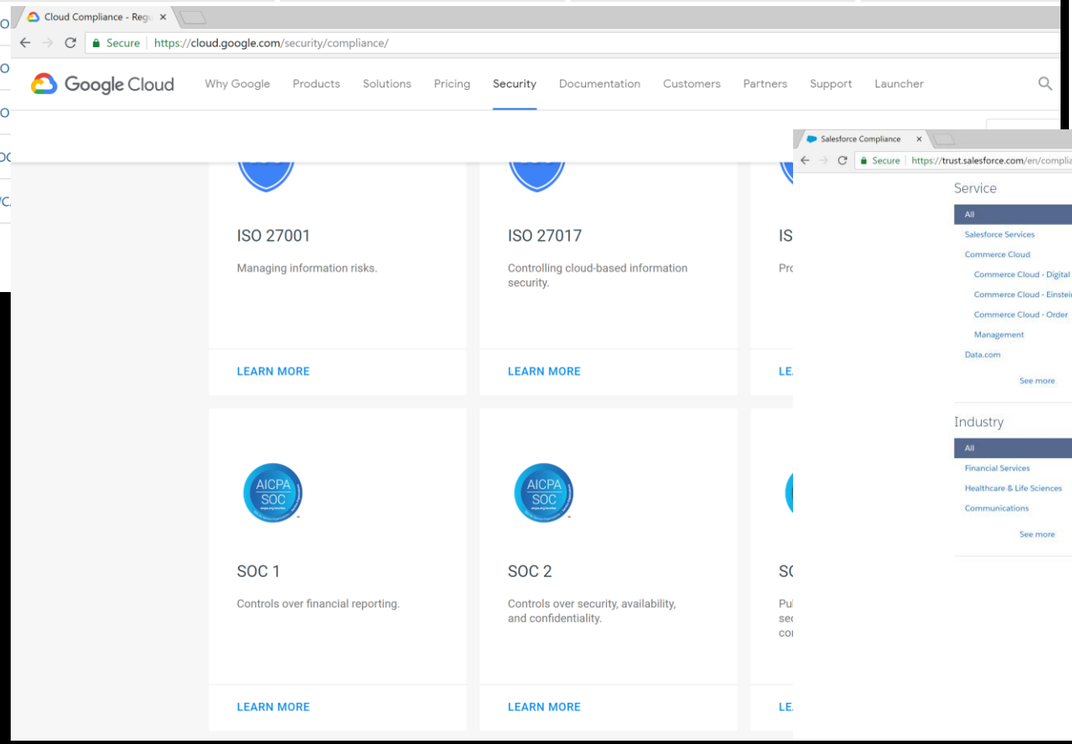
What makes them appealing?

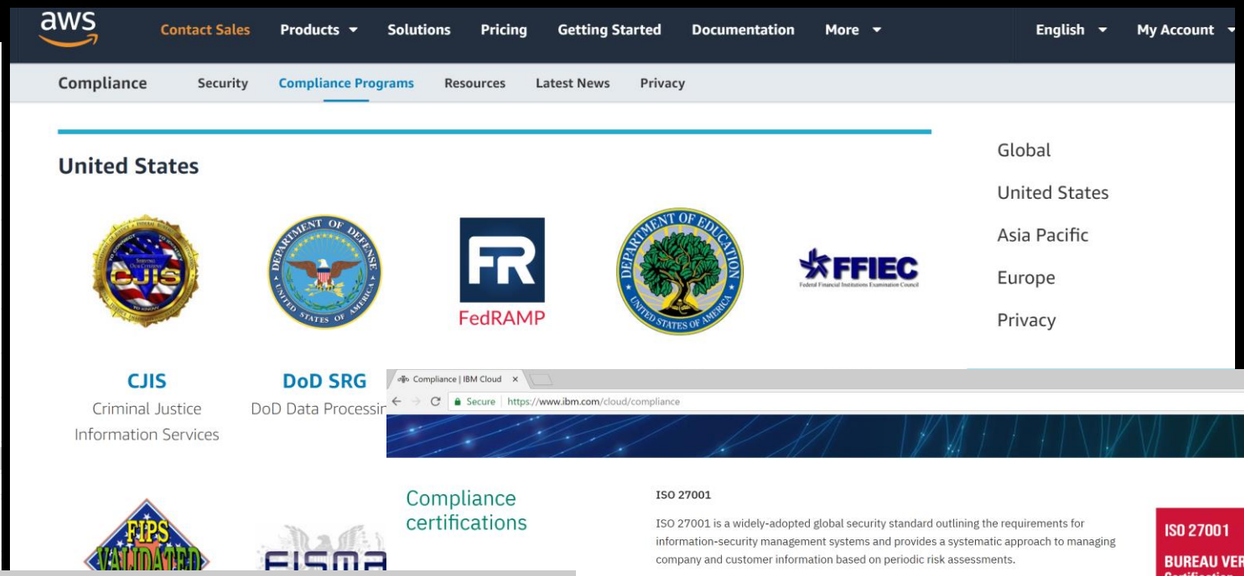
1. Comprehensive
2. Concrete
3. Constituent-Testing

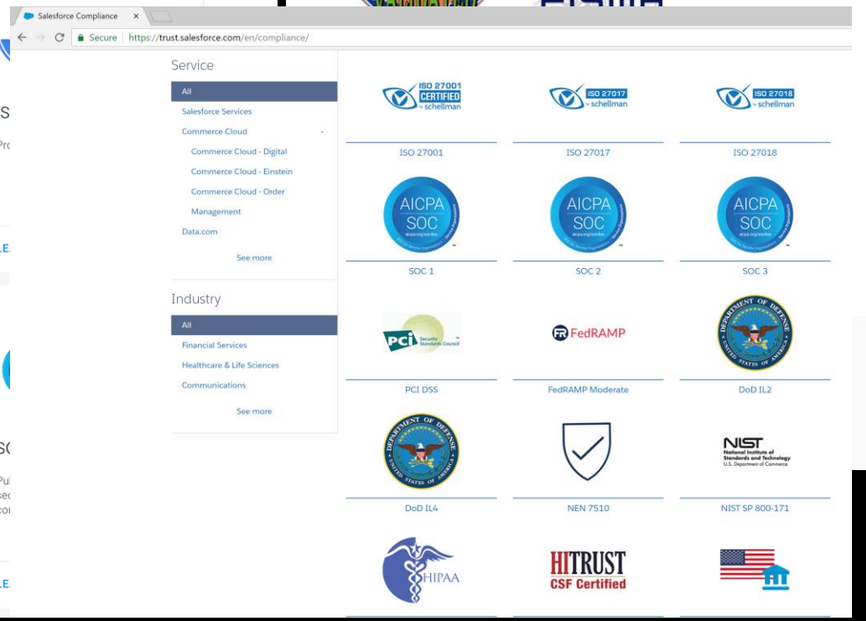
Source: https://marks.ul.com/about/ul-listing-and-classification-marks/appearance-and-significance/marks-for-north-america/

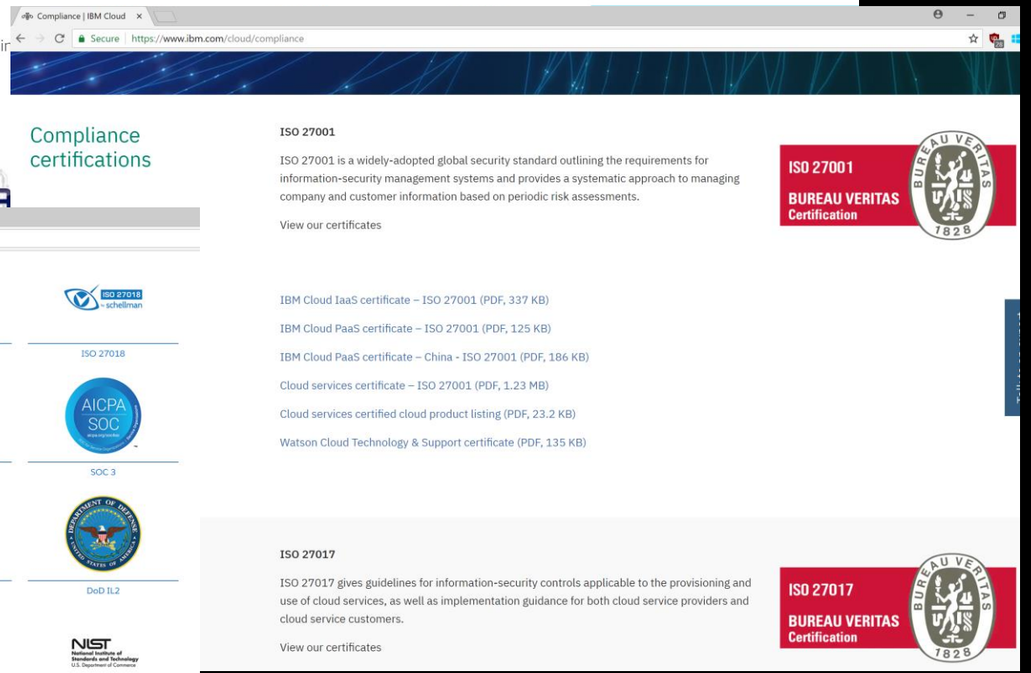
Standards and Certifications – Cloud Security

Global	US Government	Industry	Regional
CSA-STAR-Attestation	CJIS	23 NYCRR Part 500	BIR 2012 (Netherlands)
CSA-Star-Certification	DoD DISA L2, L4, L5	APRA (Australia)	C5 (Germany)
CSA-STAR-Self-Assessment	DoE 10 CFR Part 810	CDSA	CCSL/IRAP (Australia)
DFARS	EAR (US Export Administration Regulations)	CFTC 1.31	CS Gold Mark (Japan)
ISO 20000-1:2011	FDA CFR Title 21 Part 11	DPP (UK)	Cyber Essentials Plus (UK)
ISO 22301	FedRAMP	FACT (UK)	DJCP (China)
ISO 27001	FERPA	FCA (UK)	EN 301 549 (EU)



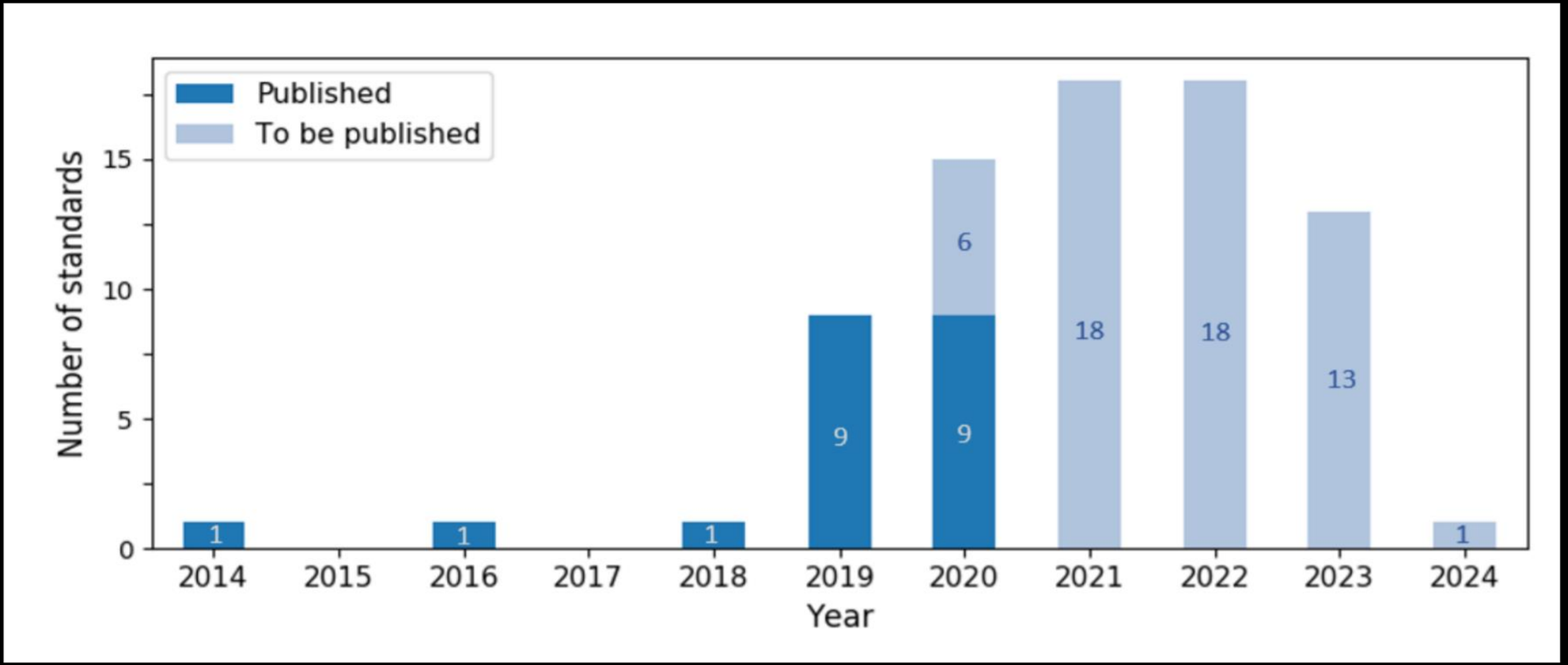






ramk@microsoft.com

Explosion in AI Standards



Source: Nativi, S. and De Nigris, S., AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework, EUR 30772 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-40325-8, doi:10.2760/376602, JRC125952.

can we calibrate trust in AI systems via standards and certifications?

@ram_ssk

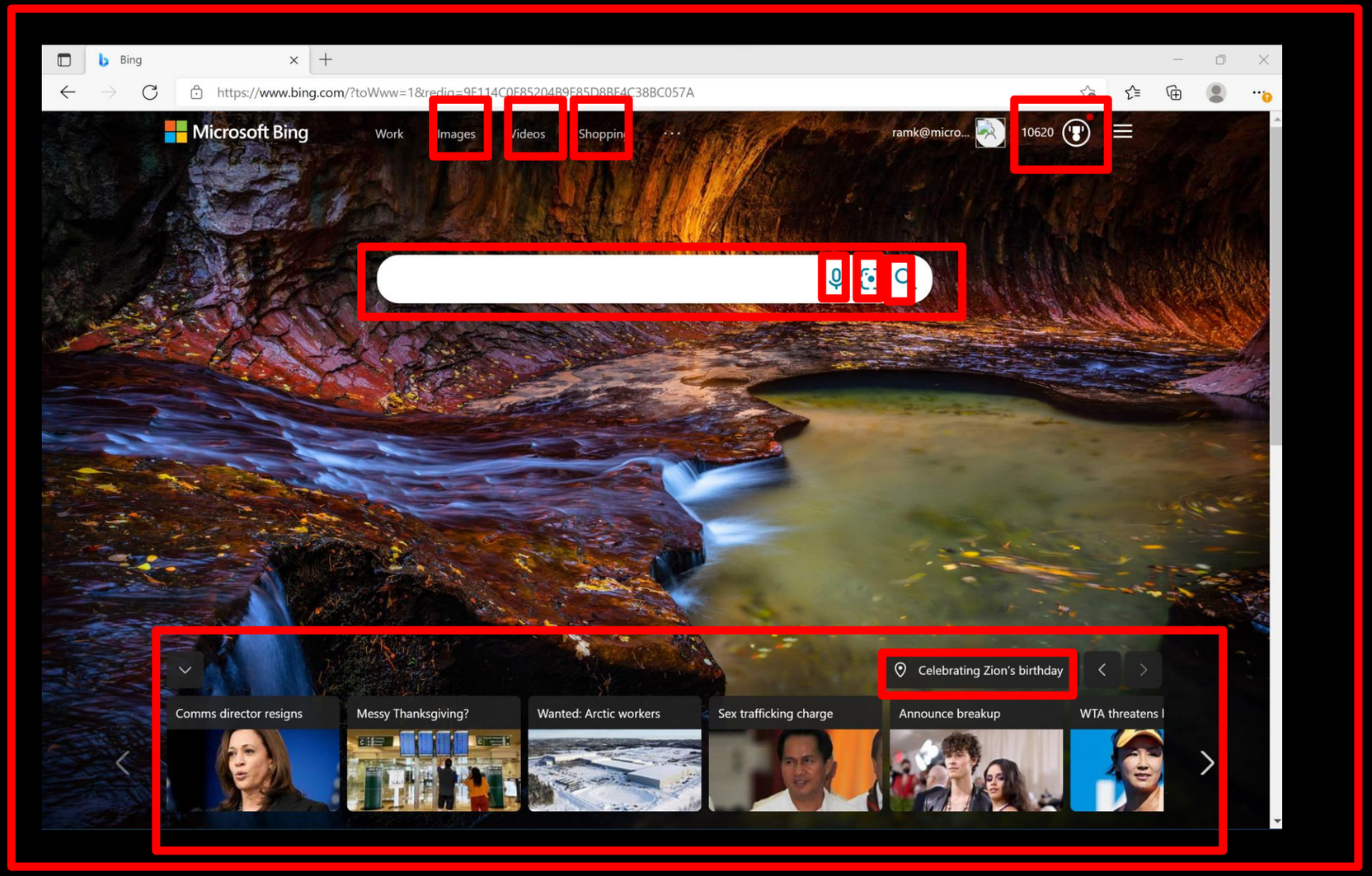
ramk@microsoft.com

can we calibrate trust in **AI systems** via standards and certifications?

@ram_ssk

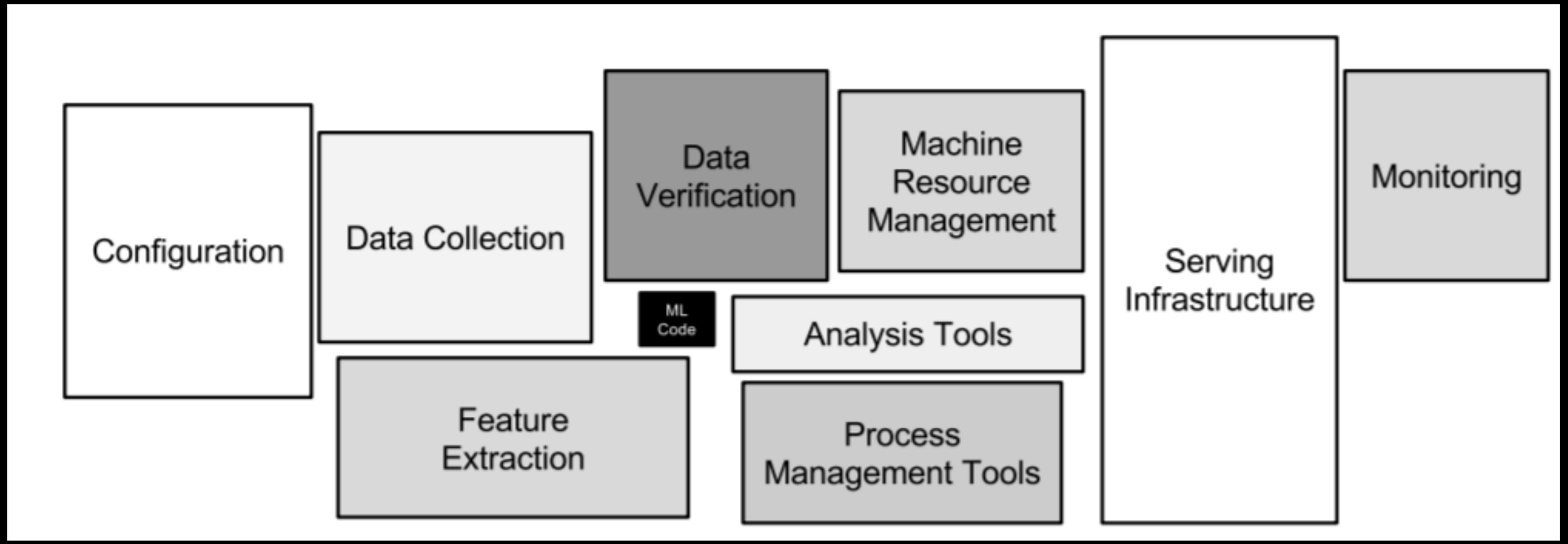
ramk@microsoft.com

AI system from Standards perspective



@ram_ssk

ramk@microsoft.com



Sculley, David, et al. "Machine learning: The high interest credit card of technical debt." (2014).

“Cool, I understand poisoning – tell me what libraries I can use to fix it”

ramk@microsoft.com

can we calibrate “trust” in AI systems via standards and certifications?

@ram_ssk

ramk@microsoft.com

“trust”

‘The Framework aims to foster the development of innovative approaches to address characteristics of trustworthiness including accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of unintended and/or harmful bias, as well as of harmful use’

Source: https://www.federalregister.gov/documents/2021/07/29/2021-16176/artificial-intelligence-risk-management-framework

ramk@microsoft.com

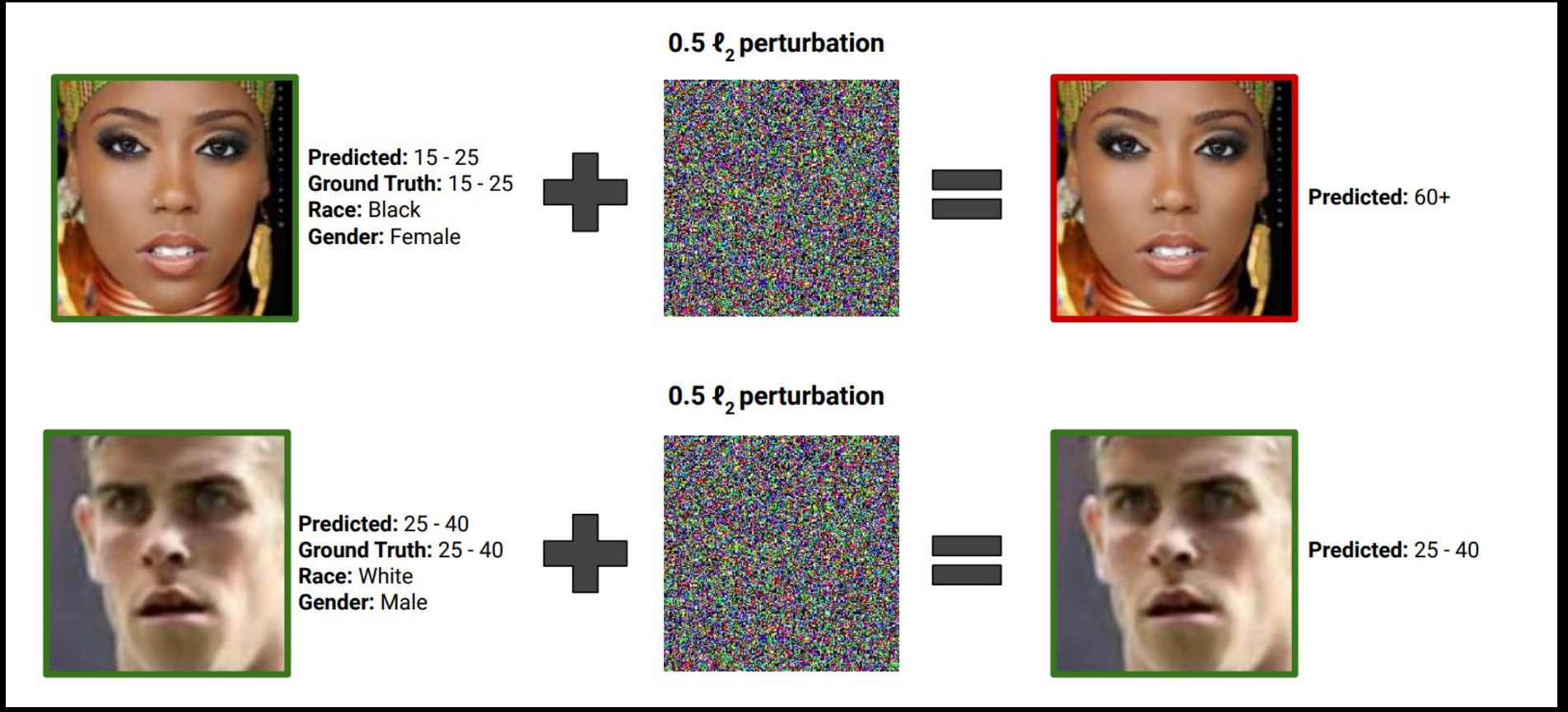
“trust”

‘The Framework aims to foster the development of innovative approaches to address characteristics of trustworthiness including accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of unintended and/or harmful bias, as well as of harmful use’

Source: https://www.federalregister.gov/documents/2021/07/29/2021-16176/artificial-intelligence-risk-management-framework

ramk@microsoft.com

Attacks do not affect all data points equally



Source :Nanda, Vedant, et al. "Fairness through robustness: Investigating robustness disparity in deep learning." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

Defenses do not protect all classes equally

67%

17%

Source: Xu, Han, et al. "To be robust or to be fair: Towards fairness in adversarial training." International Conference on Machine Learning. PMLR, 2021.

Robust or Fair?

Source:
Xu, Han, et al. "To be robust or to be fair: Towards fairness in adversarial training." International Conference on Machine Learning. PMLR, 2021.

Robust or Fair?

Robust or Explainable?

Robust or Private?

Robust or Accurate?

Robust or Safe?

Source:

Xu, Han, et al. "To be robust or to be fair: Towards fairness in adversarial training." International Conference on Machine Learning. PMLR, 2021.
Slack, Dylan, et al. "How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods." (2019).
Mejia, Felipe A., et al. "Robust or private? adversarial training makes models more vulnerable to privacy attacks." arXiv preprint arXiv:1906.06449 (2019).
Yang, Yao-Yuan, et al. "A closer look at accuracy vs. robustness." arXiv preprint arXiv:2003.02460 (2020).
Gilmer, Justin, et al. "Motivating the rules of the game for adversarial example research." arXiv preprint arXiv:1807.06732 (2018).

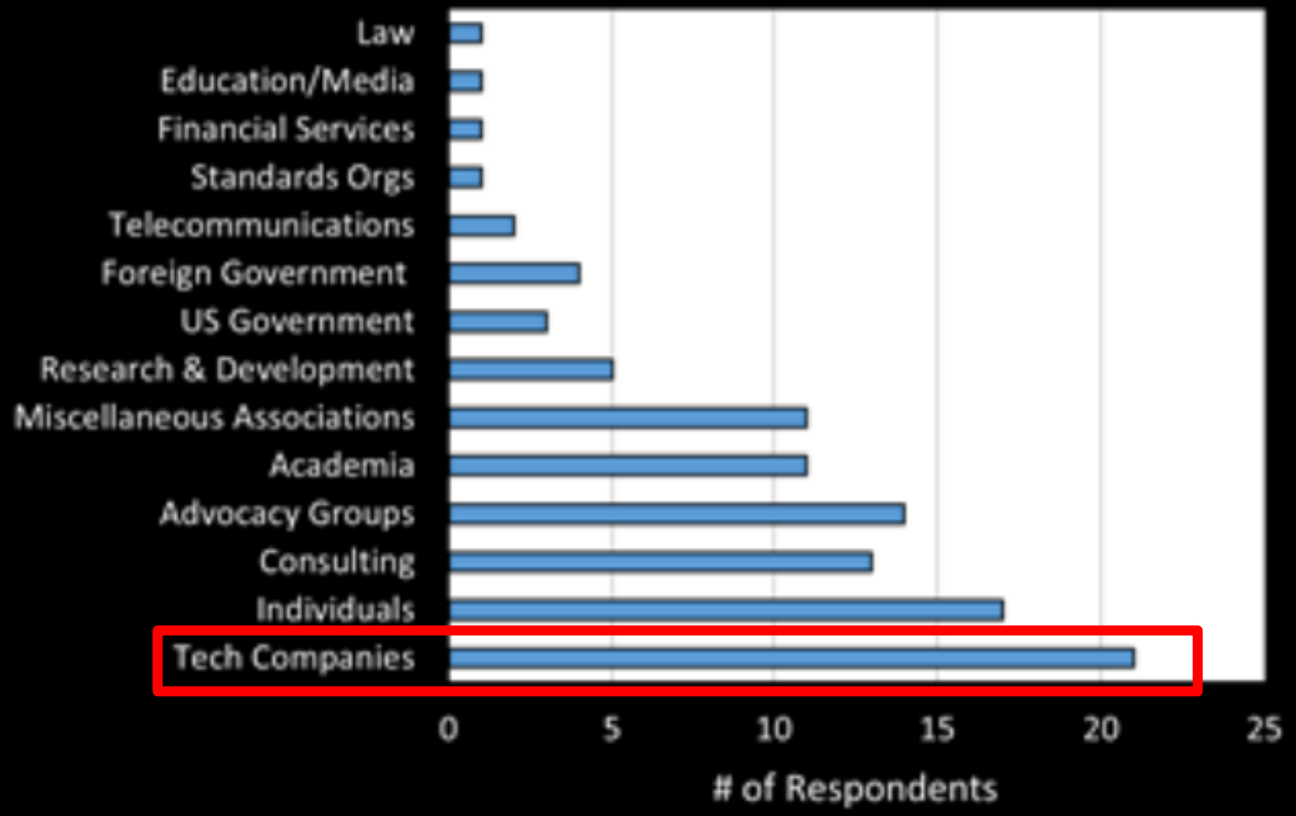
@ram_ssk
ramk@microsoft.com

can we calibrate “trust” in AI systems via standards and certifications?

@ram_ssk

ramk@microsoft.com

Whom do Standards serve?



ramk@microsoft.com

Big picture

@ram_ssk

ramk@microsoft.com

Big picture:

You simply have to trust that the very people in charge of building AI systems are adhering to the letter and spirit of these standards.

ramk@microsoft.com

Big Picture:

If we want real change, it needs to start with organization culture

ramk@microsoft.com

Inside Tesla as Elon Musk Pushed an Unflinching Vision for Self-Driving Cars

The automaker may have undermined safety in designing its Autopilot driver-assistance system to fit its chief executive's vision, former employees say.

“Where I get concerned is the language that’s used to describe the capabilities of the vehicle” - Jennifer Homendy, chairwoman of the National Transportation Safety Board,

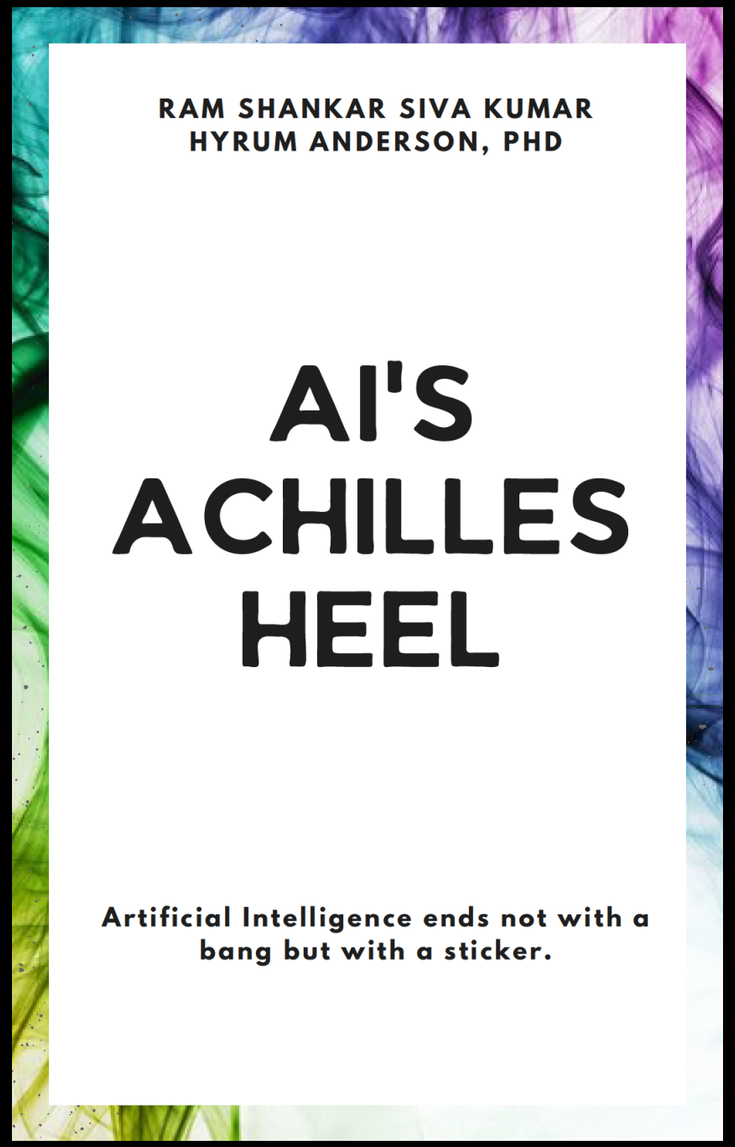
ramk@microsoft.com

Call to Action

1. We are early

2. We have muscle memory

   Standards in the software industry aren’t about proving correctness, but rather about demonstrating effort has been made to ensure correctness and best practices



www.aiachillesheel.com

ramk@microsoft.com

@ram_ssk