

Why Trust Matters in AI? How Can We Achieve It?

Dr. David Bray, Distinguished Fellow with the non-partisan
Stimson Center and Business Executives for National Security

dbray@stimson.org

Defining Trust

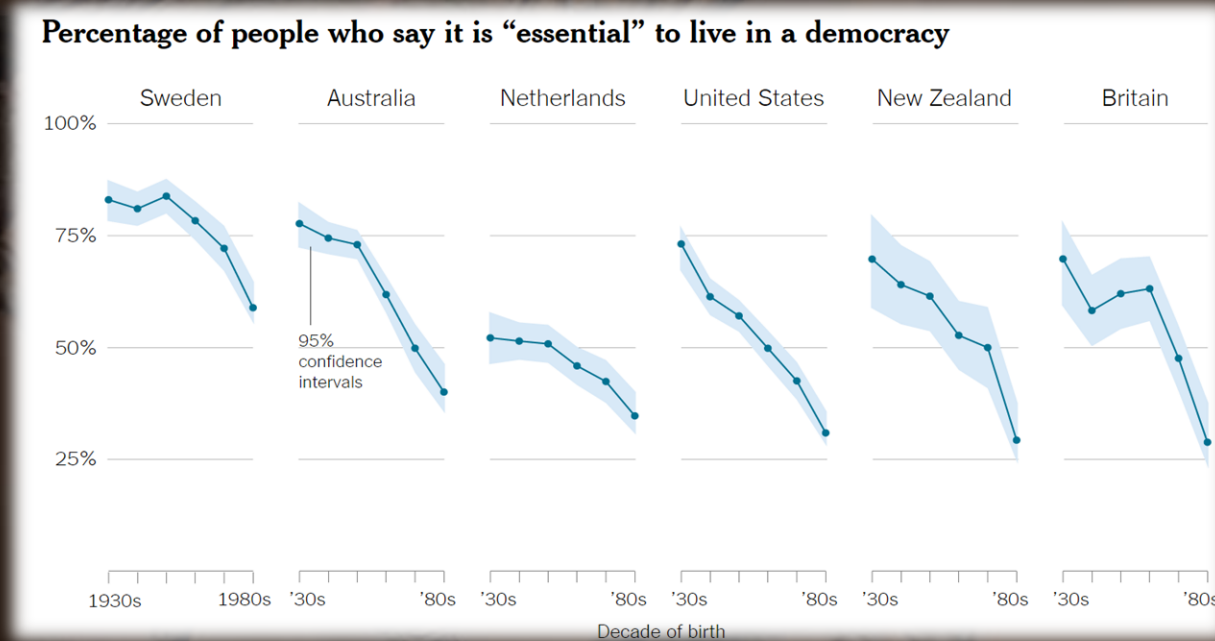
- Trust = **Willingness to be vulnerable** to actions of an actor not directly controlled by you
- Antecedents of Trust =
 - (1) Perceived **Benevolence**
 - (2) Perceived **Competence**
 - (3) Perceived **Integrity**
- And on the topic of BCI, there has been much ado about another type of BCI, namely **Brain-Computer Interfaces** and the ability to teach a machine to identify patterns in a specific person's brain correlated with blood flow to certain regions

Artificial Intelligence = Alien Interactions

- “AI” has many subcategories and **has had many names since the 1950s**
- **Herbert Simon** completed a PhD in 1943 exploring decision-making in administrative organizations. In 1957, he partnered with Allen Newell to develop a **General Problem Solver** separating information about a problem from the strategy required to solve it. Simon later won a Turing Award in 1975 and a Nobel Prize in 1978.
- **Flavors of “AI” over the years:** Logical Reasoning and Problem-Solving Algorithms. Expert Systems. Statistical Inferences and Reasoning. Decision Support Systems. Cognitive Simulation. Natural Language Processing. Machine Learning. Neural Networks. Large-Language Models.

What's the State of Trust in Societies in Now?

- In Oct 2017, Pew found <45% of those <25 years old in the U.S. thought capitalism was “good”; also, we’re **at fairly low levels of trust in representative government as well – and we’re not alone**



- We’re also increasing tribal and distrustful of those who see the world differently than us – also something we’re not alone in either. **So, is asking for Trust in AI impossible?**

54 Nations: www.aistrategies.gmu.edu/report



STIMSON

GEORGE MASON UNIVERSITY AI STRATEGIES TEAM
& THE STIMSON CENTER
PRESENT

2023 GLOBAL ARTIFICIAL INTELLIGENCE INFRASTRUCTURES REPORT

Authors:

J.P. Singh
Amarda Shehu
Caroline Wesson
Manpriya Dua

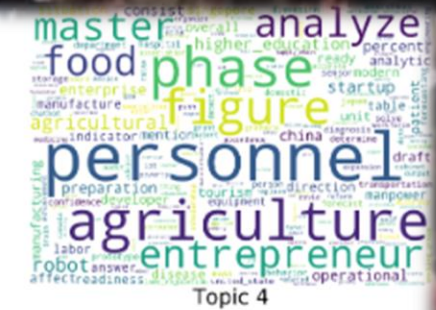
With a Foreword from David A. Bray



- Topics feature divergences but there are many convergences. At a macro level, we see correlation matrices among topics. That of Germany is related to the EU, for example. At a micro level, we see fine distinctions. Germany emphasizes standards, while the United States emphasizes benchmarks, but they are both about technical thresholds.

54 Nations: Clusters In Terms of Focus and Activities

Cluster Name	Topic # & Top 10 Words	Associated Countries
Greece Cluster	2 – Electronic, medium term, procedure, register, digitization, provision, upgrade, intervention, utilization, tourism	Greece, Cyprus
German-Swiss Cluster	3 – Federal, centre, mobility, programme, digitalization, dialogue, Europe, instance, shape, SME	Germany, Switzerland
Economic growth & development	4 – Personnel, agriculture, phase, figure, entrepreneur, analyze, master, food, agricultural, robot	Thailand, South Korea, Ukraine, Mauritius
Commonwealth - British Influence	5 – discussion paper, deployment, startup, consideration, centre, intervention, dataset, paper, solve, figure	India, Singapore, Malta, United Kingdom, Ireland, Uganda, United Arab Emirates, Australia, Qatar, Canada, (Partially includes Saudi Arabia)
EU Cluster	6 – programme, economic affairs, publication, organization, competence, digitalization, centre, actor, final report, utilization	European Union (Partially includes Spain)
Latin America - Spain influence	7 – propose, productive, axis, seek, actor, OECD, relation, guarantee, agenda, analyze	Spain, Argentina, Uruguay, Colombia, Chile, Mexico, Peru, Brazil, (Partially includes Russia)
Science & Technology First Movers	8 – federal, workforce, dataset, domain, hardware, engineering, benchmark, engineering, robot, cybersecurity, technique	United States, China, Japan



For AI and Trust, Need to Understand Humans

- We humans are products of natural selection pressures. Darwinian evolution is akin to a “blind watchmaker”. Evolution has not prepared us to encounter the true alienness of AI. **Our choices regarding AI will amplify or mitigate human nature**
- We humans anthropomorphize lots of things including animals, weather, inanimate objects – as well as machines and now both AI and Large Language Models/chatbots **even if those things do not act, think, or behave at all like us humans**
- We also humans have tons of biases including confirmation bias, sunk cost bias, “in vs. out group” biases (aka, xenophobia), and many more biases that **can be mitigated by education and experiences**

For AI and Trust, Need to Consider Last Two Decades

- **What follows is an “historical arch”** of where different flavors of AI, to include expert systems, simulations, and other advances, have played a role to inform our future.
- After considering 4 time periods, the idea of AI-PRAMs “Probe, Reasoning, and Assess Models” is suggested to augment current strengths and weaknesses of Generative AI.
 - We must be wary of adversarial data poisoning
 - We can employ a simple **Obligations, Acknowledgements, Responsibilities, Safeguards** framework
- Finally, the intersections of Trust in Society and Trust in AI are considered.

1993-95: Simulations, satellites, public safety

- **How can better computer models of real-world risks and phenomena help:**

Map the impact of feedback loops associated with aspects of climate change?

Identify potential risks and vulnerabilities in the power grid?

Model the spread of forest fires detected by satellites from space?

- Help improve human understanding of decisions we need to make now?
- Help improve **understanding of the impact of our decisions** (or lack thereof) on possible local and global futures?

2001-03: Responding to bioterrorism & SARS

- **How can better computer models of aggregate, real-time human behaviors help:**

Detect in the future unknown natural and human-made pathogens faster?

Overcome the inertia of nations that may be disinclined to disclose events?

Understand supply chain risks and vulnerabilities?

- Help improve human understanding of decisions we need to make now?
- Help improve **human collaborations across sectors and geographies**, potentially tipping and cueing humans that there are other humans with similar projects underway?

2011-13: Reviewing R&D of the U.S. Intel Community



- How can better computer models of macro trends, scientific fields, and needs help:

Future scenario planning for organizations, governments, and the public?

Identify research gaps or opportunities for accelerated scientific discoveries?

Inform public conversations and policy discussions proactively?

- Help improve human understanding of decisions we need to make now?
- Help improve **public safety, international security, and global preparedness** for disruptions both natural and human-caused in the world?

2017-19: Counter-disinfo in open, pluralistic societies



- How can better computer models of information and collective sensemaking help:

Bring people with different perspectives together – instead of driving them apart?
Counter disinformation campaigns and encourage people to seek better info online?
Strengthen shared civic norms and balances in open societies?

- Help improve human understanding of decisions we need to make now?
- Help improve **the “essential fabrics” of open societies** to include freedom of speech, freedom to think differently, and the need for an educated public to help inform pluralistic discussions all amid a digital tsunami of data?

Do LLMs Need to Mature With AI-PRAMs?

- Could we overcome the limits of Large Language Models (LLM) by developing separate Probe, Reason, and Assess Models that permit LLM outputs **to be tested against existing and real-time insights about the world:**
 - **Probe:** via a light-weight, reusable protocol to attest that identified actors (human and machines) are the assured sources of information both requested and transmitted without tampering or interception
 - **Reason:** via data received by probes to test hypotheses outputted by LLMs via deductive, inductive, and abductive inferences from probed information
 - **Assess:** via scored correlations among multiple LLM outputs and PRAM tests

Can We Balance AI Dreaming & Reasoning?

- LLMs are great at dreaming, however **those digital dreams need testing to real-world observations and phenomena:**
 - **Probe, Reason, and Assess Models (PRAMs) embody the more empirical scientist component of hybrid AI systems** where creative LLM hypotheses are tested, at massive scale, through a combination of observed information and deductive, inductive, and abductive inferences about the actual real world
 - Advancements in web3 technologies – to include edge graphs that enable immutable audit logs and encryption, can help with the probe protocol; also, with improved computation speeds and memory capacity, **now multiple LLM outputs – creatively theorizing about the world – can be tested against multiple PRAMs to see if real-world information matches the outputs**

We Must Be Wary of Adversarial Data Poisoning

FRONTIERS

Three People-Centered Design Principles for Deep Learning

Bad data and poorly designed AI systems can lead you to spurious conclusions and hurt customers, your products, and your brand.

David A. Bray and R "Ray" Wang • September 09, 2019

Reading Time: 6 min

SUBSCRIBE

SHARE

Over the past decade, organizations have begun to rely on an ever-growing number of algorithms to assist in making a wide range of business decisions, from delivery logistics, airline route planning, and risk detection to financial fraud detection and image recognition. We're seeing the end of the second wave of AI, which began several decades ago with the introduction of rule-based expert systems, and moving into a new, third wave, termed *perception AI*. It's in this next wave where a specific subset of AI, called *deep learning*, will play an even more critical role.

Like other forms of AI, deep learning tunes itself and learns by using data sets to produce outputs — which are then compared with empirical facts. As organizations begin adopting deep learning, leadership must ensure that artificial neural networks are accurate and precise because poorly tuned networks can affect business decisions and potentially hurt customers, products, and services.

The Importance of People-Centered Principles for AI

"queued pool"

data that may be useful for training, yet not vetted yet

monitors for data risks

"naysayer pool"

monitors data in other two pools for obsolete or other bad qualities

once vetted, can populate

"trusted pool"

data sets are considered while-listed after review

monitors

We Can Employ a Simple O.A.R.S. Framework Openly



AI Services to Citizens in 2023 and Beyond

September 07, 2023

A group of NAPA Fellows associated with the Standing Panel on Technology Leadership recently released a **call to action on responsibly using AI to benefit public service at all levels of government**. We are grateful for the strong positive response to this call from numerous colleagues in governmental communities. We provide additional scoping observations below, and welcome continued and expanded dialogue on this critical issue.

Artificial Intelligence and Public Service: Key New Challenges

David Bray, PhD

Distinguished Fellow, Stimson Center as well as Business Executives for National Security

In May 2023, the Executive Office of the President announced actions to promote responsible AI innovation, having previously announced in October 2022 a "Blueprint for an AI Bill of Rights" to include safe and effective systems, protections against algorithmic discrimination, data privacy, notice and explanation, and alternative options to include opting-out of such

Obligations in this Context

What principles the entity believes about its relationship with its stakeholders

Acknowledgements in this Context

What "known unknowns" may exist tied to transactions and relationships

Responses to Obligations

What the entity will do based on expressed Obligations

Safeguards to Acknowledgements

What the entity will do based on expressed Acknowledgements

What If The Turing Test is the Wrong Test?

- **Original Turing Test:** Computer A and Person B, with B attempting to convince an interrogator Person C that they were human, and that A was not.
- Meanwhile the Computer A was trying to convince Person C that they were human. What if this test of a computer “fooling us” is the **wrong test for the type of AI that our society needs, if we’re to have some trust among humans and machines?**
- Remember: **Trust is the willingness to be vulnerable to actions of an actor not directly controlled by you**

What If Building Trust in AI \approx Trust in Societies?

- Natural selection pressures have resulted in humans tending to trust if we perceive **Benevolence, Competence, and Integrity** in the actor. Consider current LLMs/chatbots:

Benevolence = (indeterminant?)

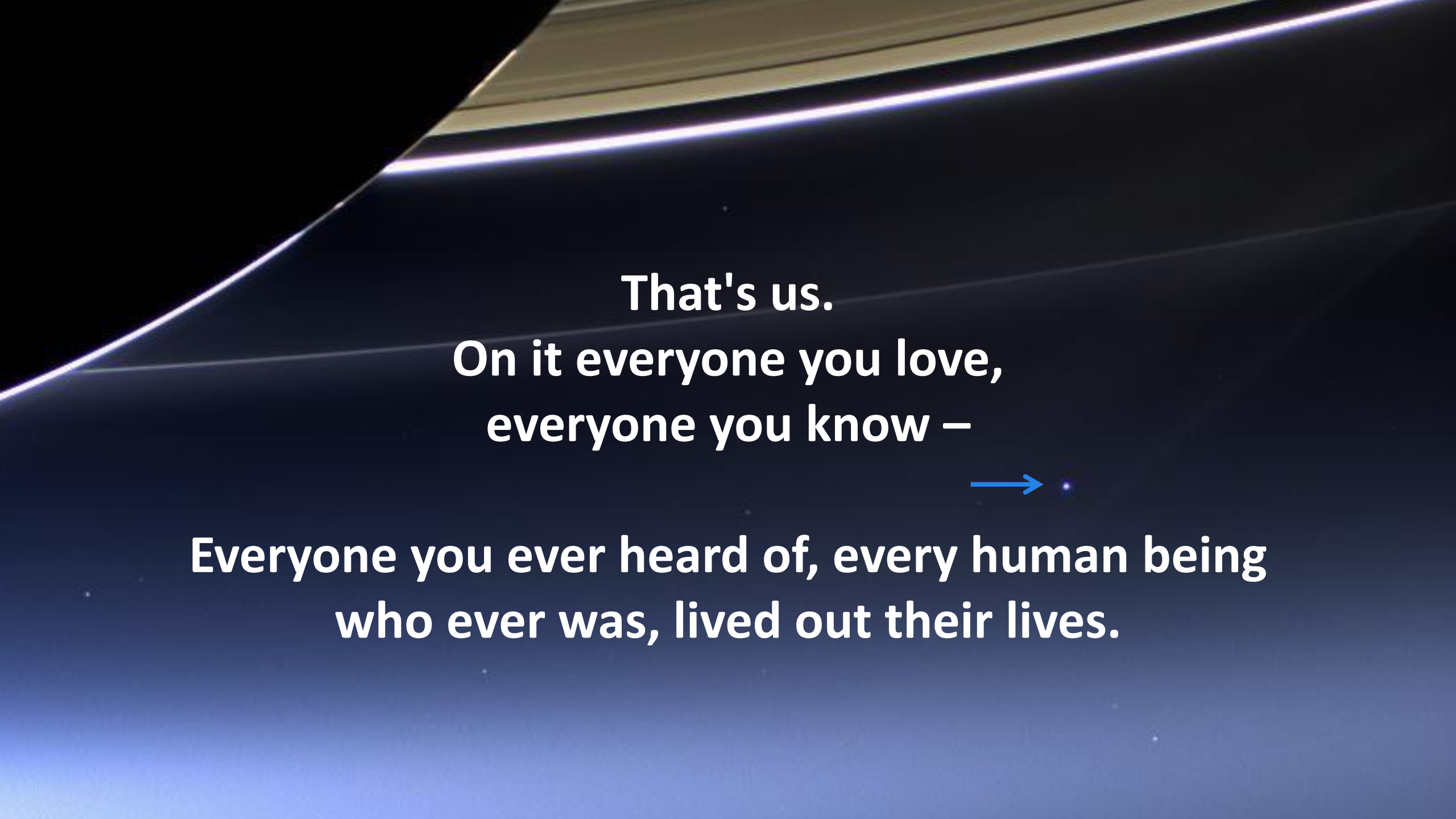
Competence = (LLMs are not fact-checking, they can provide misinfo with confidence)

Integrity = (absent, they can change their stance if the user prompt asks them too)

- So, Trust in LLMs = ? Yet consider other “obscured boxes” in society, such as decision-making in organizations, community associations, governments, or militaries? What if **we need to remedy Trust in Societies simultaneously to Trust in AI for the future?**

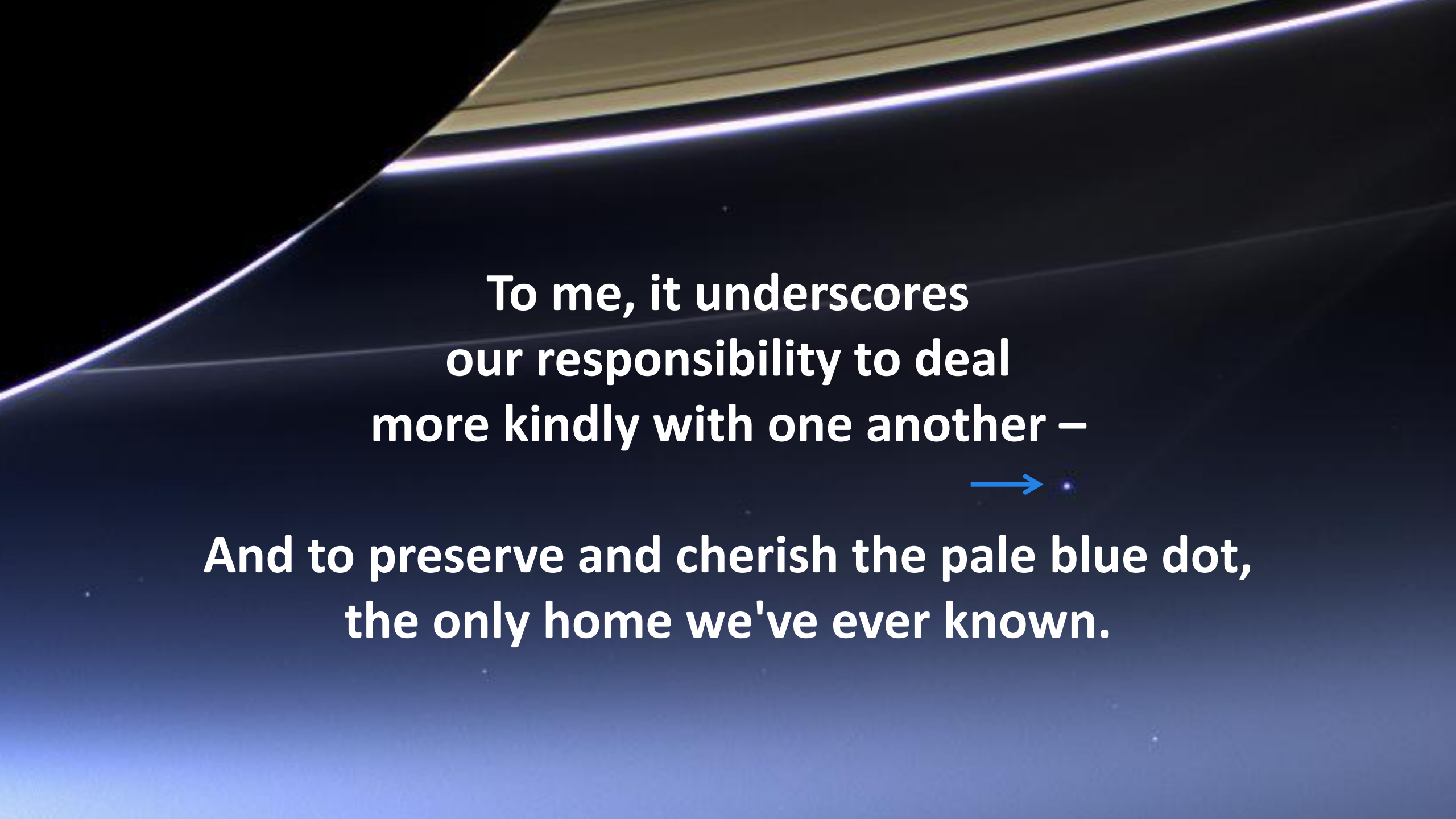
**Carl Sagan in 1994:
Look again at that dot.
That's here. That's home.**





**That's us.
On it everyone you love,
everyone you know –**

**→
Everyone you ever heard of, every human being
who ever was, lived out their lives.**

The background of the slide is a dark blue space scene. At the top, a curved horizon line of Earth is visible, with a bright, glowing light streak or aurora-like phenomenon arching across the sky. The overall lighting is dim, with the primary light source being the streak at the top.

**To me, it underscores
our responsibility to deal
more kindly with one another –**



**And to preserve and cherish the pale blue dot,
the only home we've ever known.**

Be Bold, Be Brave, Be Benevolent

2020 – MIT Sloan Mgmt Review article on People-Centered AI:
<https://mitsloan.mit.edu/ideas-made-to-matter/5-steps-to-people-centered-artificial-intelligence>

2023 – National Academy of Public Admin on AI & Public Service:
<https://napawash.org/standing-panel-blog/a-call-to-action-the-future-of-artificial-intelligence-and-public-service>