# Is Your AI QUALITY Good Enough To Save Humanity?

# AI QUALITY

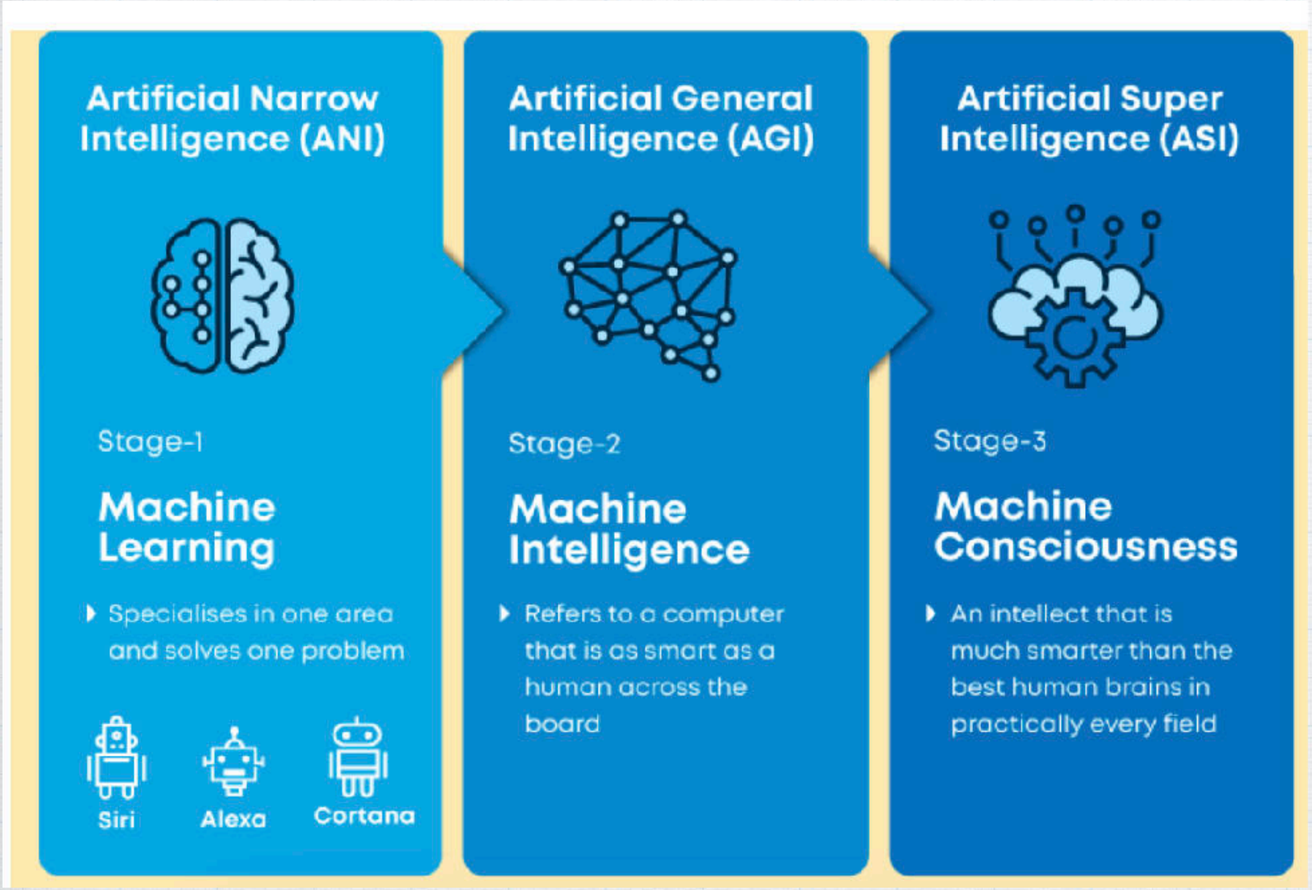
TOM GILB
BCS 3 July 2024

45 minutes talk and 15 minutes Q&A
Recorded (1 MIN./SLIDE)

tom@Gilb.com

Gilb International (ret.)
Norway

These slides folder (+ appendix)
is-your-ai-quality-good-enough-to-save-humanity
SLIDES FOLDER : https://www.dropbox.com/scl/fo/u1fisxhd5cyrp7iij2i98/AAVYSATm4Vt40X0t8lEJdpQ?rlkey=3n4nn6077nig6gvo5qiuu0em8&dl=0




https://www.mygreatlearning.com/blog/what-is-artificial-intelligence/

SLIDES FOLDER

DOI to slides pdf: https://doi.org/10.13140/RG.2.2.10935.12960

# Outline: 'AI QUALITY'

1. Current ways to 'understand' AI systems
2. Why these current AI Models are defective.
3. Multidimensional AI Models for Qualities, And Costs
4. The AI Stakeholder Model
5. Understanding AI Strategies in light of Multiple Quality and Cost Attributes
6. Teaching The Multi-AI model to students and research students as a tools for researching AI developments.
7. The Quality distinctions between Large Language Model (**LLM**) **AI** and the next **AI** generation **(AGI, Artificial General Intelligence)**
8. The Penta Model as a basic high level view of any AI system.
9. The Principles of AI Understanding, outside the blackbox.


# "AI Safety" is a Musk Concern

# AI Safety

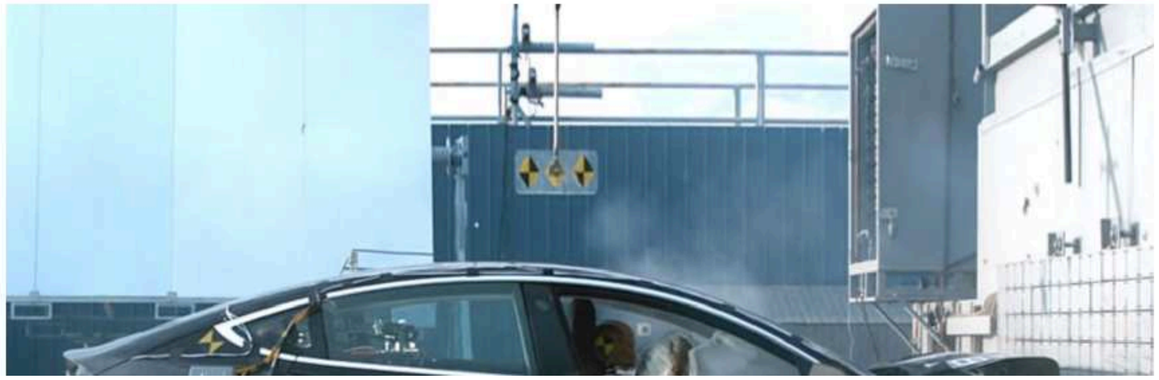
## According To Tesla, The Model 3 Is The Safest Car On The Planet

OCT 9, 2018 / BY JAY TRAUGOTT IN TESTING

### And the number two and three spots go to...

Last month it was announced the new Tesla Model 3 earned a perfect five-star crash test rating in every category from the National Highway Transportation Safety Administration. Most impressive. And now Tesla says its new mid-size sedan has the lowest probability of causing injury for any vehicle ever tested by the US government agency.

"Not only has Model 3 achieved a perfect 5-star safety rating in every category and sub-category, but NHTSA's tests also show that it has the lowest probability of injury of all cars the safety agency has ever tested," Tesla wrote in a blog post. This same post also explains why the Model 3 is so safe.



# Are you an AI Engineer, yet?

1. How many of you know how to quantify (define a metric for) AI System '**Transparency**', and can show me you have done it, in writing?

3. How many of you know how to quantify (define a metric for) AI System '**Security**', and can show me you have done it, in writing?

# 'AI QUALITY'

# Some insights into A.I. methods

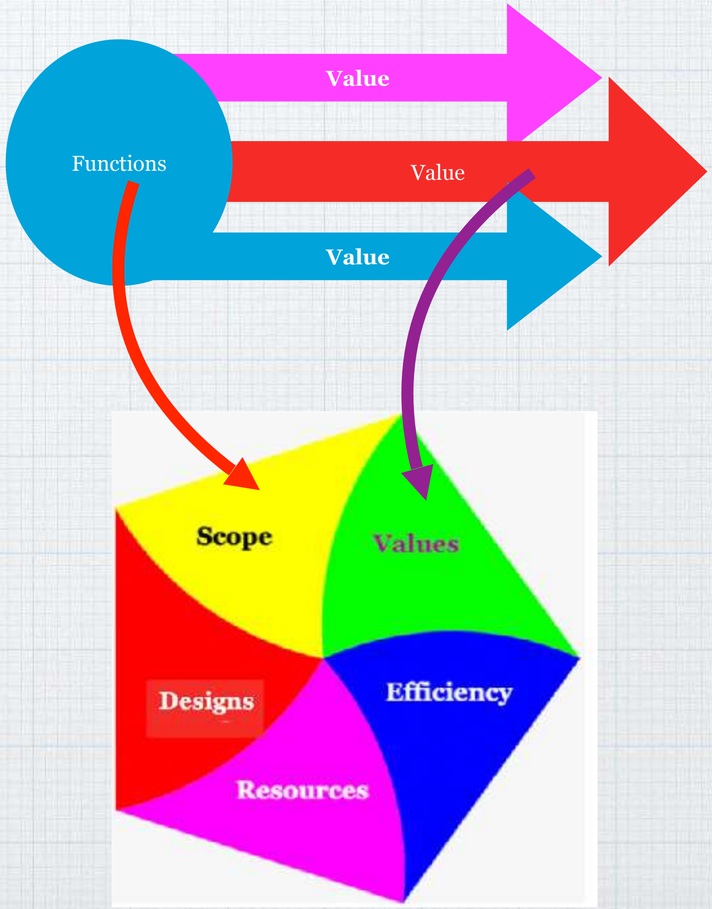
https://altern.ai/topic/resources

# 1. Current (BAD) ways to ‘understand’ AI systems

- Design (‘How’, inside the black box)
  - DESIGNS: We use Big Data, LLM (Large Language Models), ‘Data Reflectivity’
  - Functions (‘What it does’, inside black box) “It does poems and art”
- Values-Qualities (‘How Well’, a result of ‘Design’), like Security, Usability, Transparency, Traceability.
  - “It is easy to use, but not very trustworthy, and difficult to understand, why it gives such answers.”
  - Usually described in imprecise terms, non-numeric, unmeasurable (like ‘easy use’, ‘fairly secure’)
- Resources (‘Costs’, a result of ‘Design’, and ‘Operate’)
  - “It is free to play with, expensive to build, and answers quickly.

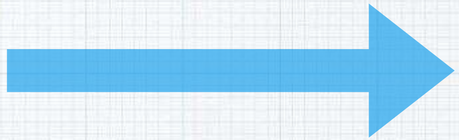
A large language model: LLM
is a language model consisting of a neural network with many parameters, trained on large quantities of unlabelled text using self-supervised learning.

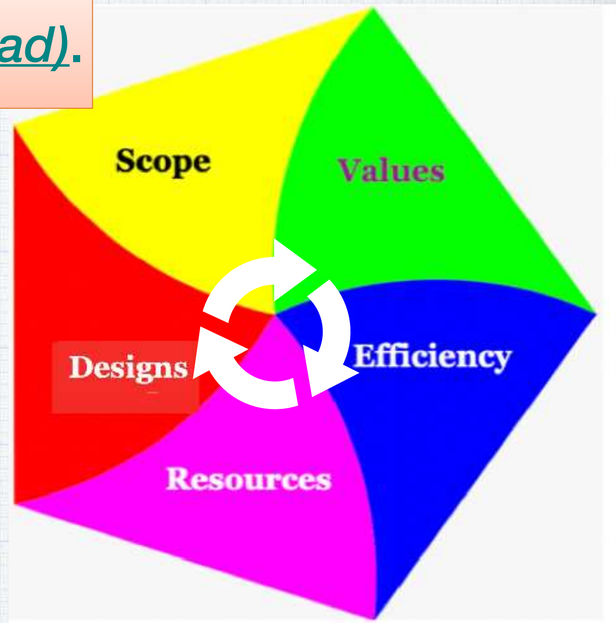
LLMs emerged around 2018

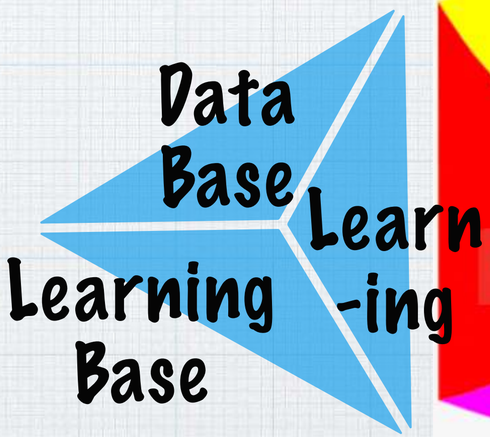


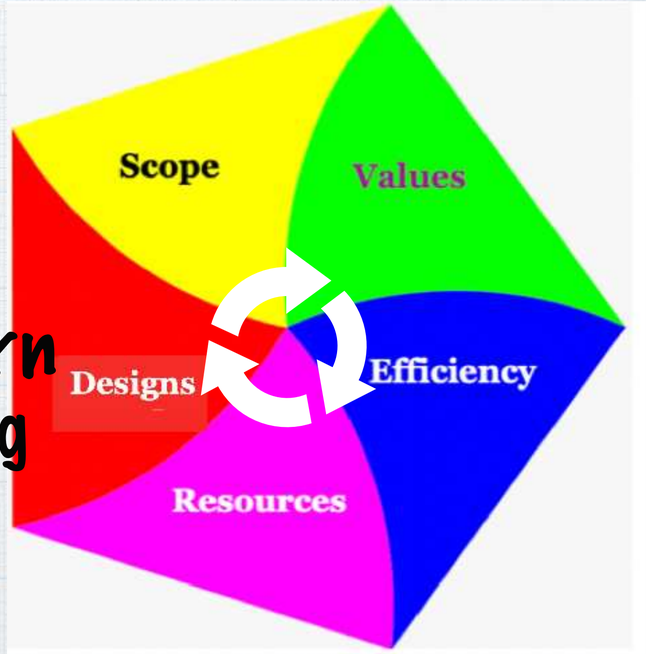
## 2.1 Why these current AI 'Models' (ways of describing an AI system) are *defective (bad)*.

- Design ('How', inside the black box)
  - Design is secret, hidden, non-transparent, un-intelligible, complex, changing (updates), expanding automatically
- Functions ('What', inside black box, part of 'Scope')
  - A result of the Design, same problems as above (un-intelligible)
  - What kind of poems? What kind of art? (Highly ambiguous)
- Values-Qualities ('How Well', a result of 'Design'), like Security, Usability, Transparency.....
  - a result of Unintelligible Design. Unstable, unpredictable
  - usually described in imprecise terms, non-numeric, unmeasurable (like 'easy use', 'fairly secure')
  - there are very many (20+) interesting qualities 'all at once'
- Resources ('Costs', a result of 'Design')
  - Same problem as Qualities, for same reasons: just nice words (but, https://ai-benchmark.com/)





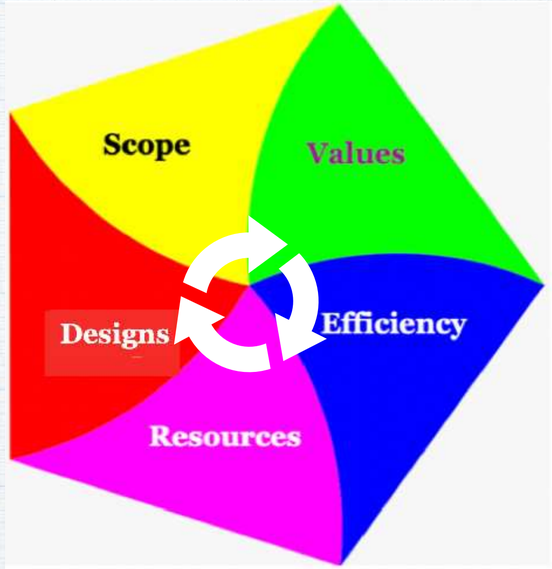

The 'Penta' Model

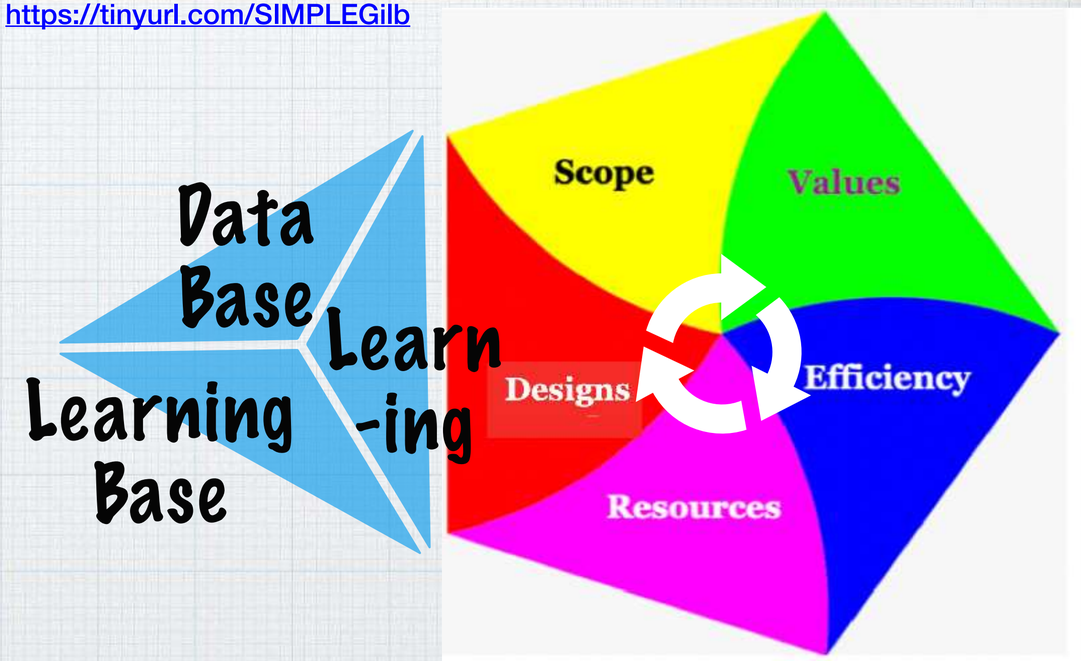
## 2.2 Why these current AI 'Models' (ways of describing an AI system) are defective.

- The technical structure of the AI system is inherently unstable, because...
- The 'conscious design' (red triangle) (the program and algorithms, by people) .....
  - can be changed by unidentified people
  - for unidentified reasons,
  - and even introduce unintended side-effects (bugs),
  - and receive insufficient quality assurance and testing (until proven otherwise).

The Penta Model Paper Alone August 2022
https://tinyurl.com/PentaPaper
also Chapter 4 of this (SIMPLE) book
https://tinyurl.com/SIMPLEGilb

Initial AI-app or code, state

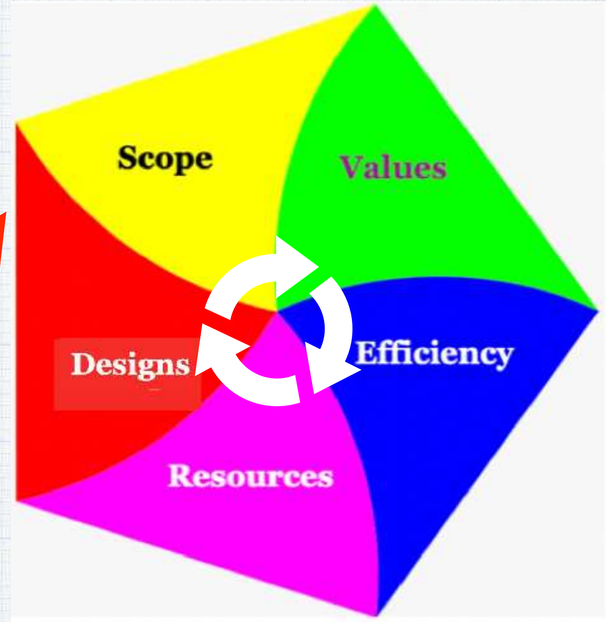


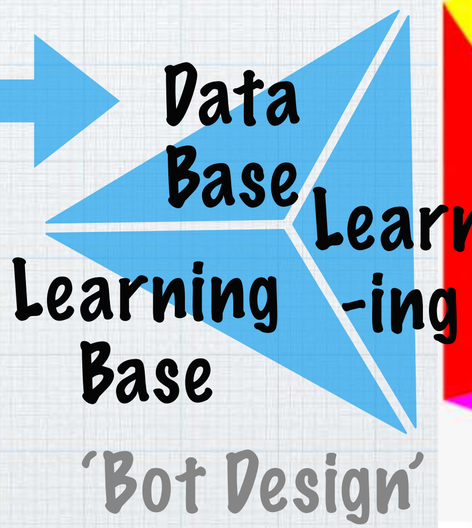



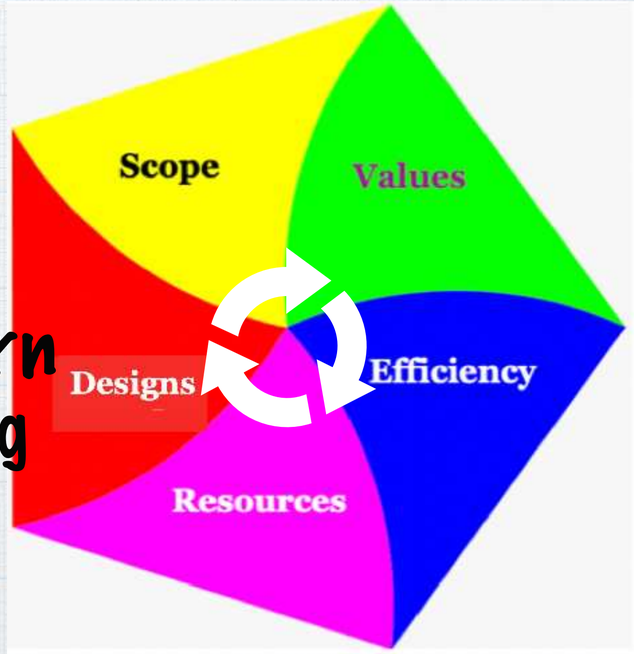

Just above is the later state of AI-app due to Training, and more data

## 2.3 Why these current AI 'Models' (ways of describing an AI system) are defective.

- The technical structure of the AI system is inherently unstable, because...
  - the 'Bot-Design' (Effectively = changes to the Human Design, via Data),
    - and enabled by the (ongoing) human re-design)
  - is done automatically, instantaneously, and by rules dictated by 'Casual Data' .....
    - = anything on internet, at any time, or generated by itself (AI), or fed by Hackers, including Fake News)



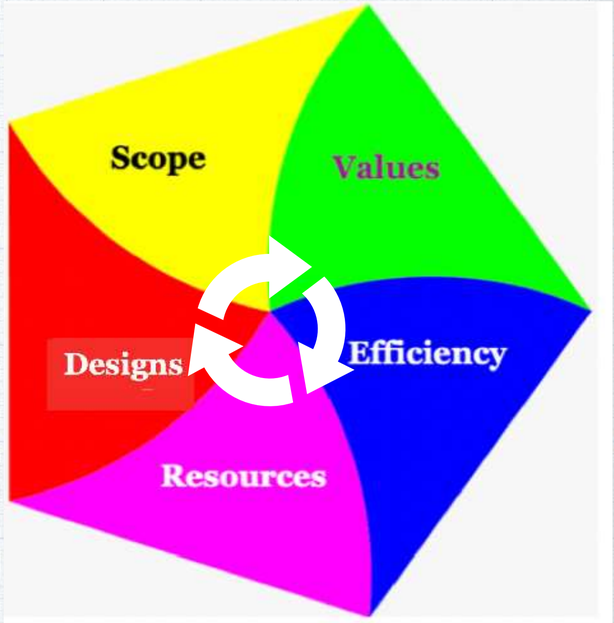



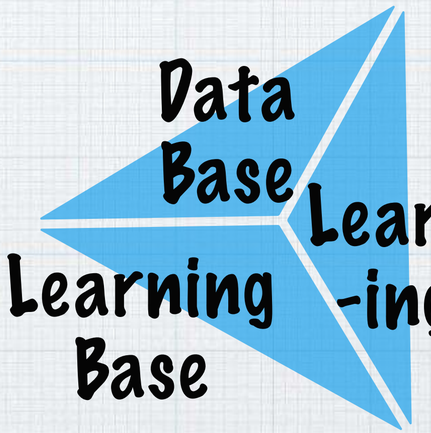



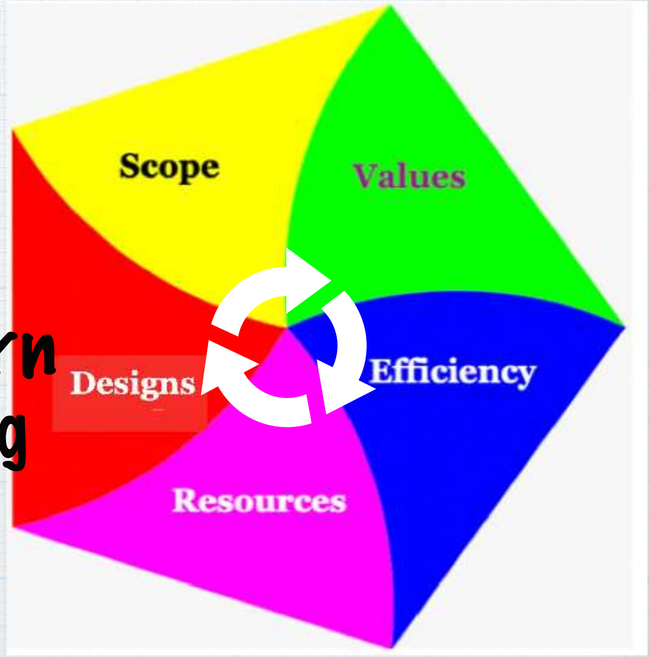

- 'Casual Data': data that 'just happens'

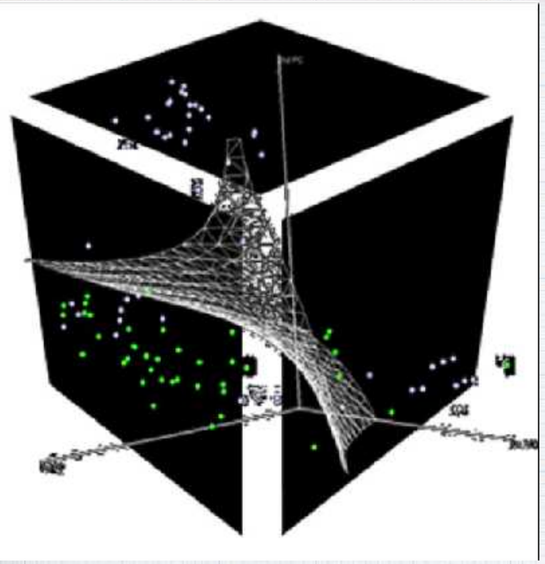
## 2.4 Why these current AI 'Models' (ways of describing an AI system) are defective.

- The technical structure of the AI system is inherently unstable, because...
- THE CONCLUSION
  - You cannot know the 'design' + casual data - that gave you some 'answer'
  - So knowing the 'design' (inside the AI Black box) is not a good way to understand the AI system
  - in order to decide if is truthful, accurate, false, fake, useful.

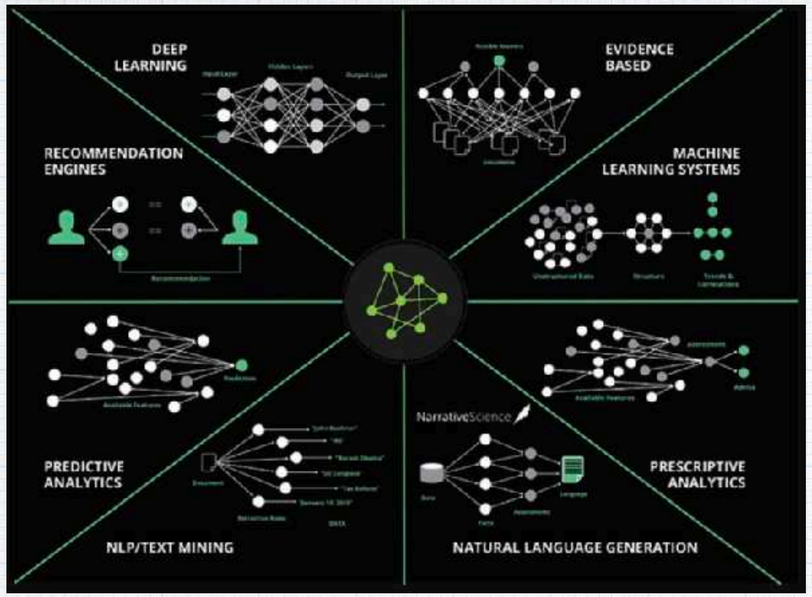


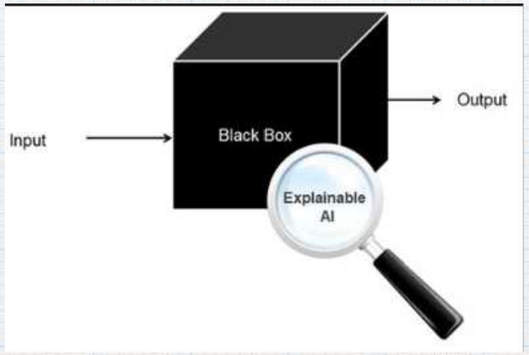


Mystical Complexity in the AI System

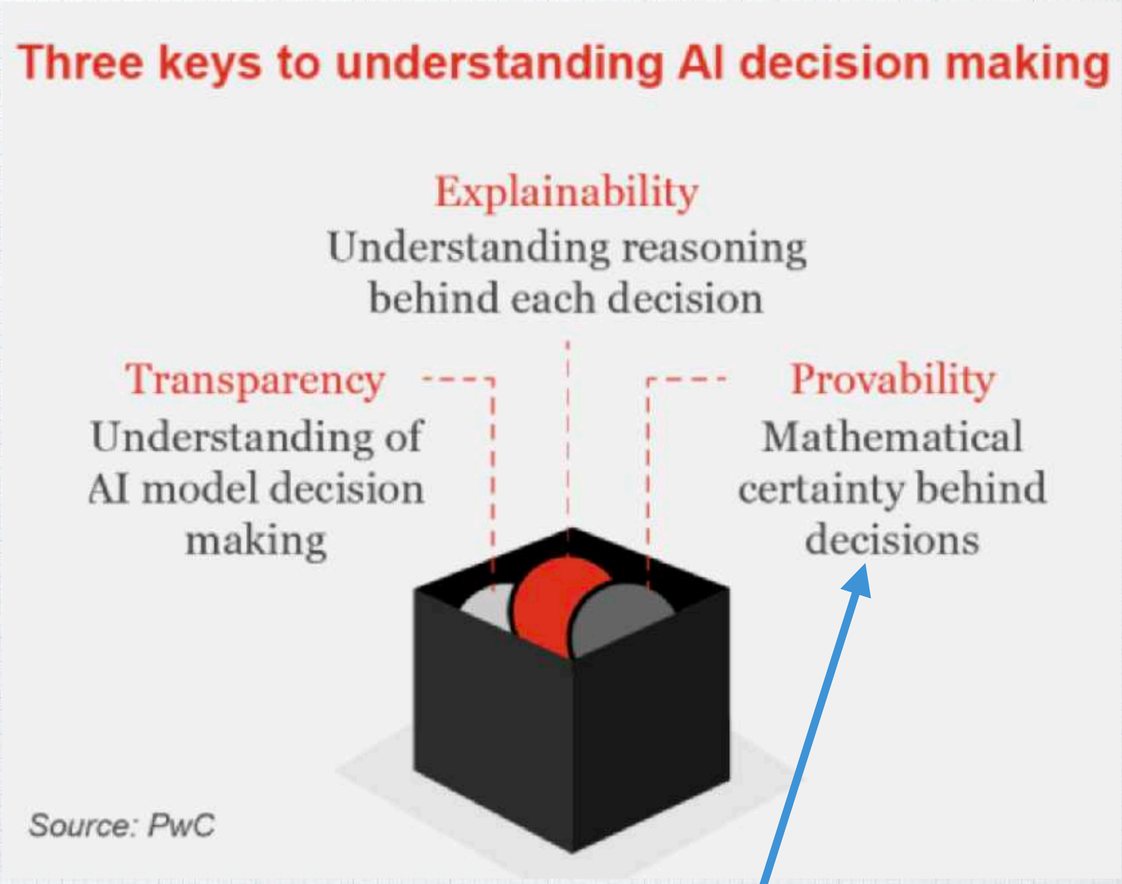



Classifying Scope of The AI System

# Understanding Critical AI Properties By quantification and measurement, of, for example, 'TRANSPARENCY'




Analyzing the Outputs from a set of Inputs. Tricky when much of the input is 'the internet'




Trying to understand the Qualities of the AI system: the effects for stakeholders.
Notice PWC: 'bla bla words.
but no hint
of 'quantification', as we present in this talk.

Sources in presenter notes

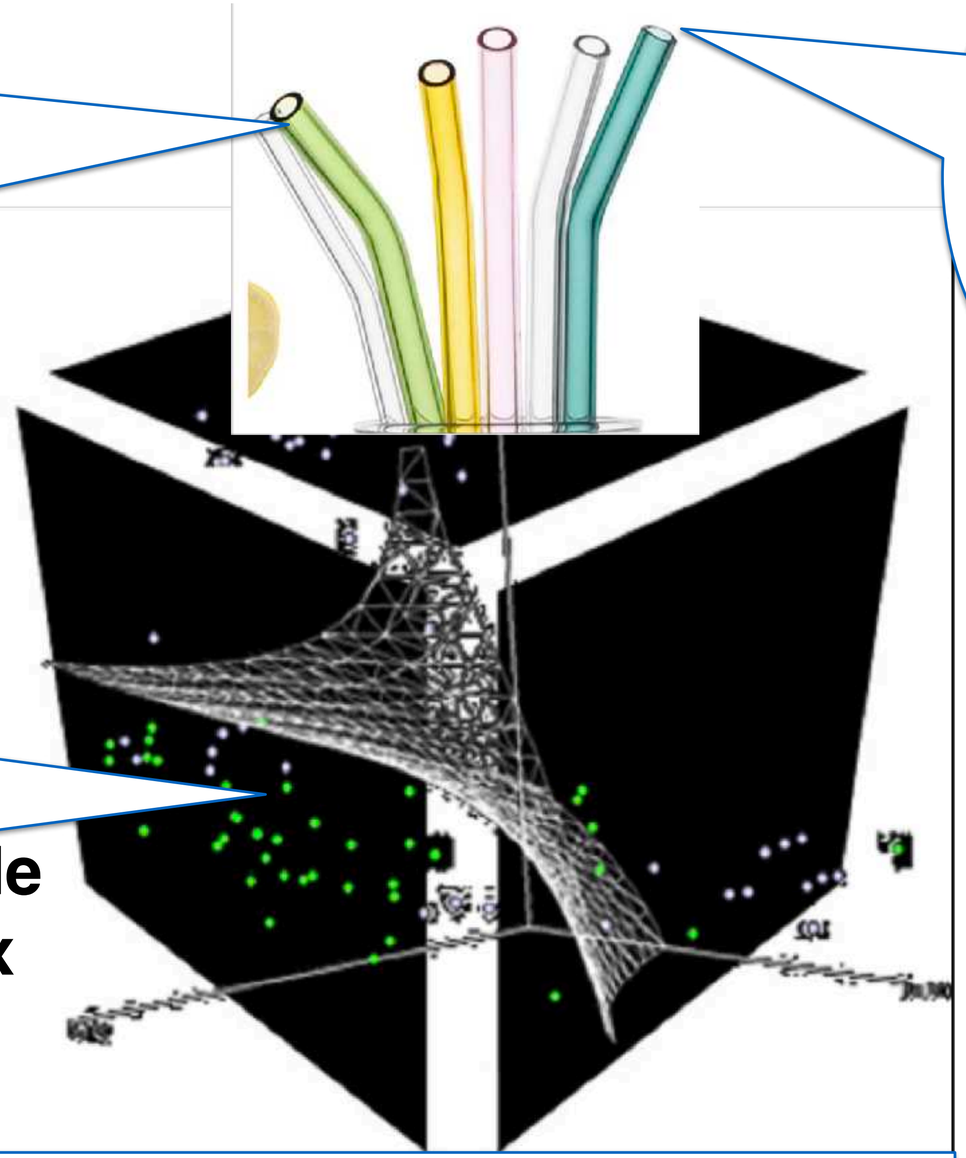
Technoscopes

10 Non-Existent ‘facts’ per report

42 Intentional Fake news per report

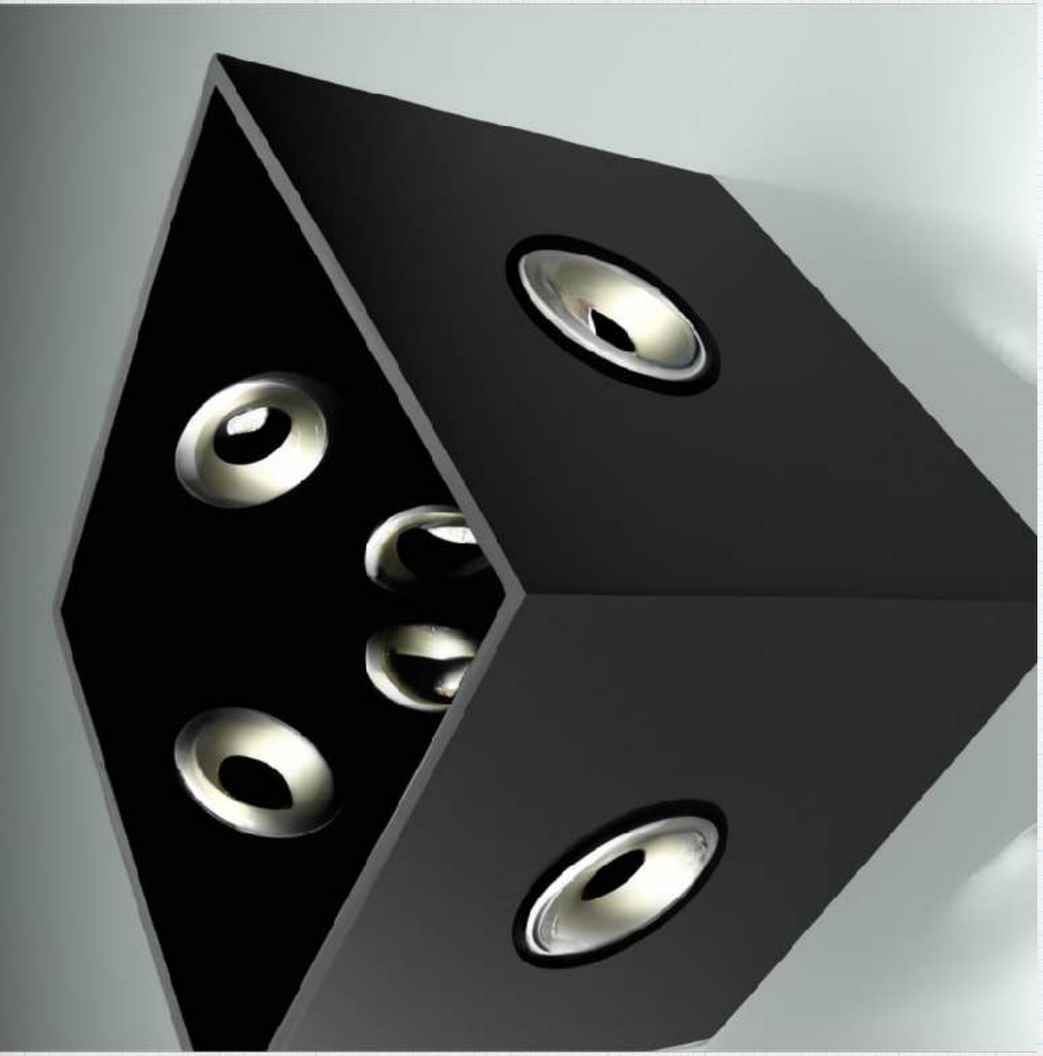
Factuality

Misleadingness



Unintelligible Complexity inside the AI Black box

Mystical Complexity in the AI System



https://app.simplified.com/tools/text-to-image/my-creations

DALL-E. https://labs.openai.com/e/NfHUMF9htol5j847yGwGQc4i/boxHJS683d7MIhs91vsu0Giz

AI TOOL PROMPT: “a black box with chaos inside
and 5 sets of 3 peepholes on outside of it
one in each side, 3d rotating”


Full interview: "Godfather of artificial intelligence" talks impact and potential of AI

CBS Mornings
2.5M subscribers
Subscribe
10K
Share

493K views 3 weeks ago #artificialintelligence #tech

https://www.youtube.com/watch?v=qpoRO378qRY

Hinton is mainly interested in Studying how the brain works.

So we need to note that He says the AI we are developing is 'very different from our intelligence'

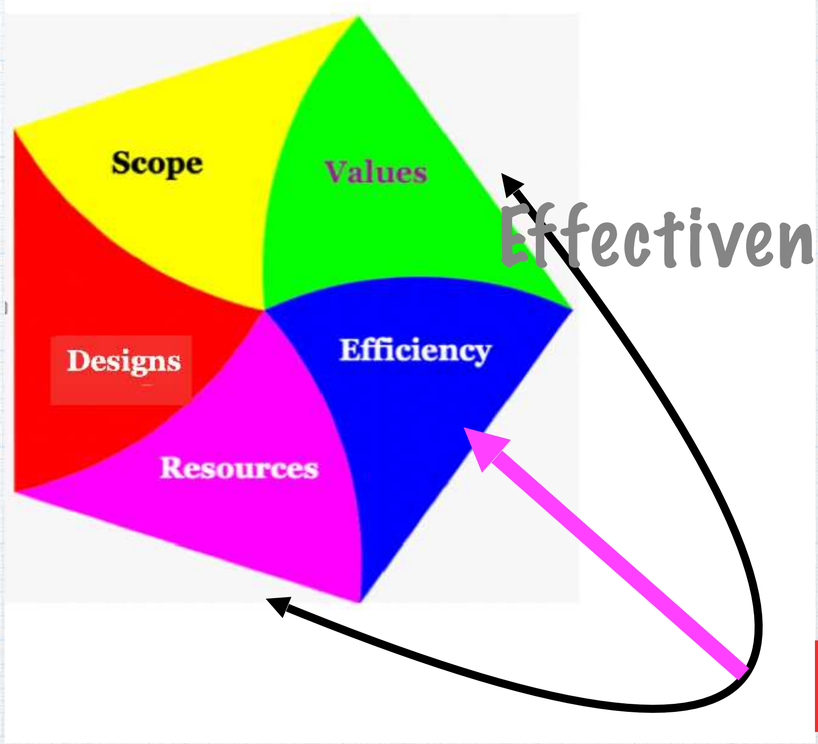
How different, in what dimensions or attributes..... ?

This is the question I am trying to find a way to answer, in this talk. TG

**Geoffrey Hinton is considered a godfather of artificial intelligence, having championed machine learning decades before it became mainstream. As chatbots like ChatGPT bring his work to widespread attention, we spoke to Hinton about the past, present and future of AI. CBS Saturday Morning's Brook Silva-Braga interviewed him at the Vector Institute in Toronto on March 1, 2023.**

# 3.1 Multidimensional AI Models for Qualities, And Costs

- So here is my advice… (. how to understand 'A.I.' )
  - (And it is good general advice for complex systems)**
- **Measure qualities and other stakeholder values**
- **Use a set of AI 'qualities', to 'understand' an AI system**, in terms of what it means for stakeholders (people)
  - The total of this qualities set defines AI 'Effectiveness'
- Use a set of AI-system **costs** to understand 'economics'
  - The total set of these costs helps us understand to AI-resources needed.
- The Values/Costs helps us understand the 'efficiency' of the AI-system




V/R = E

The Ratio of a set of Values / Set of Resources = Efficiency of the Designs

**The Penta Model Paper Alone August 2022**
**https://tinyurl.com/PentaPaper**

** FREE COPY BOOK : TECHNOSCOPES. https://www.gilb.com/offers/YYAMFQBH/

# 3.2 Multidimensional AI Models for Qualities, And Costs

* Here are some examples of AI Quality sets
  * IEEE Ethics Model
    * 1. Human Benefit
    * 2. Responsibility
    * 3. Transparency
    * 4. Education and Awareness

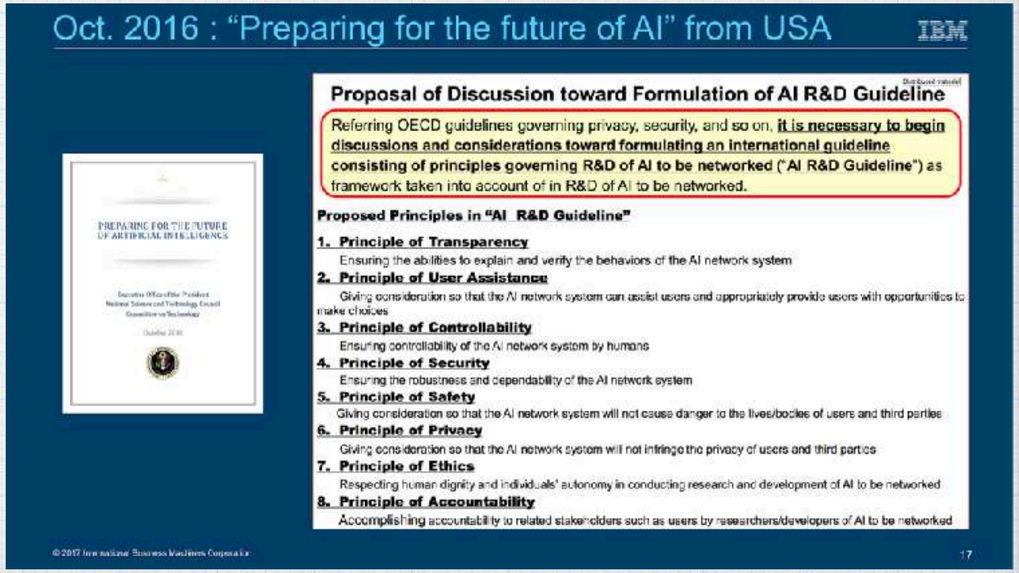
**Candidate Recommendation**

Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined. For designers, such standards will provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency. (The mechanisms by which transparency is provided will vary significantly, for instance (1) for users of care or domestic robots a why-did-you-do-that button which, when pressed, causes the robot to explain the action it just took, (2) for validation or certification agencies the algorithms underlying the AI/AS and how they have been verified, (3) for accident investigators, secure storage of sensor and internal state data, comparable to a flight data recorder or black box.)

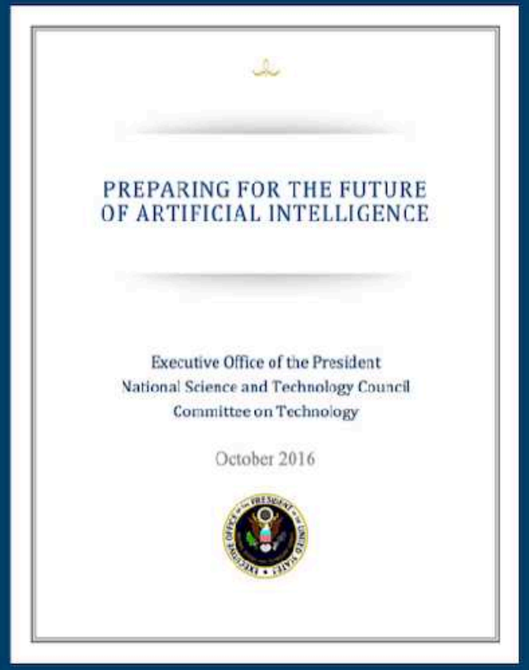
    * https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_general_principles.pdf
  * AI for The White House
  *

XAI Talk 2019 Slides

I Quantified, in later slides



# Oct. 2016 : “Preparing for the future of AI” from USA


PREPARING FOR THE FUTURE
OF ARTIFICIAL INTELLIGENCE

Executive Office of the President
National Science and Technology Council
Committee on Technology

October 2016

Distributed material

## Proposal of Discussion toward Formulation of AI R&D Guideline

Referring OECD guidelines governing privacy, security, and so on, **it is necessary to begin discussions and considerations toward formulating an international guideline consisting of principles governing R&D of AI to be networked** (“**AI R&D Guideline**”) as framework taken into account of in R&D of AI to be networked.

**Proposed Principles in “AI R&D Guideline”**

1. **Principle of Transparency**
   Ensuring the abilities to explain and verify the behaviors of the AI network system
2. **Principle of User Assistance**
   Giving consideration so that the AI network system can assist users and appropriately provide users with opportunities to make choices
3. **Principle of Controllability**
   Ensuring controllability of the AI network system by humans
4. **Principle of Security**
   Ensuring the robustness and dependability of the AI network system
5. **Principle of Safety**
   Giving consideration so that the AI network system will not cause danger to the lives/bodies of users and third parties
6. **Principle of Privacy**
   Giving consideration so that the AI network system will not infringe the privacy of users and third parties
7. **Principle of Ethics**
   Respecting human dignity and individuals’ autonomy in conducting research and development of AI to be networked
8. **Principle of Accountability**
   Accomplishing accountability to related stakeholders such as users by researchers/developers of AI to be networked

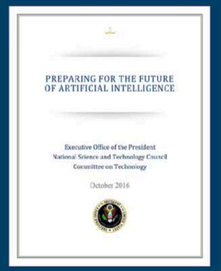
# "SERIOUSLY guys, don't try to bullshit a bullshitter!"

- "Ensuring the abilities to explain and verify the AI Network System"
  - 7 ambiguous words
  - "Ensuring":
    - improving the probability, (2 ambiguities here)
    - guaranteeing,
    - making sure it is perfectly explained,
    - making sure it is perfectly effective
    - And 500 more

Oct IBM

Proposal of Discussion toward Formulation of AI R&D Guideline

Referring OECD guidelines governing privacy, security, and so on, **it is necessary to begin discussions and considerations toward formulating an international guideline consisting of principles governing R&D of AI to be networked** ("**AI R&D Guideline**") as framework taken into account of in R&D of AI to be networked.




**Proposed Principles in "AI R&D Guideline"**

1. **Principle of Transparency**
   Ensuring the abilities to explain and verify the behaviors of the AI network system
2. **Principle of User Assistance**
   Giving consideration so that the AI network system can assist users and appropriately provide users with opportunities to make choices
3. **Principle of Controllability**
   Ensuring controllability of the AI network system by humans
4. **Principle of Security**
   Ensuring the robustness and dependability of the AI network system
5. **Principle of Safety**
   Giving consideration so that the AI network system will not cause danger to the lives/bodies of users and third parties
6. **Principle of Privacy**
   Giving consideration so that the AI network system will not infringe the privacy of users and third parties
7. **Principle of Ethics**
   Respecting human dignity and individuals' autonomy in conducting research and development of AI to be networked
8. **Principle of Accountability**
   Accomplishing accountability to related stakeholders such as users by researchers/developers of AI to be networked



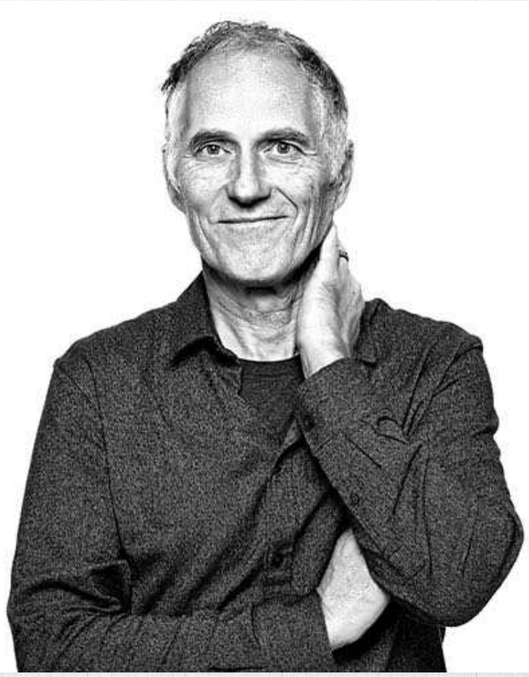
17

# …we need to design and report on the metrics ("You can't regulate what you don't understand.")

https://www.oreilly.com/content/you-cant-regulate-what-you-dont-understand-2/

*But whose human values? Those of the benevolent idealists that most AI critics aspire to be? Those of a public company bound to put shareholder value ahead of customers, suppliers, and society as a whole? Those of criminals or rogue states bent on causing harm to others? Those of someone well meaning who, like Aladdin, expresses an ill-considered wish to an all-powerful AI genie?*

*There is no simple way to solve the alignment problem. But alignment will be impossible without robust institutions for disclosure and auditing.* ***If we want prosocial outcomes, we need to design and report on the metrics that explicitly aim for those outcomes and measure the extent to which they have been achieved. That is a crucial first step, and we should take it immediately.*** *These systems are still very much under human control. For now, at least, they do what they are told, and when the results don't match expectations, their training is quickly improved. What we need to know is what they are being told.*



https://www.oreilly.com/tim/
tim@oreilly.com
tim@oreilly.com

Thanks for this Pawel Nowak

## 3.3 Multidimensional AI Models for Qualities (quantified), And Costs

* Here are some examples of AI Quality sets
  * IEEE Ethics
    * 1. Human Benefit
    * 2. Responsibility
    * **3. Transparency**
    * 4. Education and Awareness

* https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_general_principles.pdf

Candidate Recommendation

**Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined.**

For designers, such standards will provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency. (The mechanisms by which transparency is provided will vary significantly, for instance (1) for users of care or domestic robots a why-did-you-do-that button which, when pressed, causes the robot to explain the action it took, (2) for validation or certification agencies the algorithms underlying the AI/AS and how they have been verified,

(3) for accident investigators, secure storage of sensor and internal state data, comparable to a flight data recorder or black box.)

IEEE Recommends

# 3.4 Multidimensional AI Models for Qualities (quantified), And Costs

## Candidate Recommendation

**Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined.**

**Principle 4 — Transparency**

**Level:** Organization , **Status:** Not Determined **Type:** Value, **Labels:** no labels Edit

**Is Part Of:** GENERAL IEEE PRINCIPLES

**Consists Of:** Explicability Interpretability Traceability

Show Sidebar

"Transparency" One-Pager Objective

Status 24 Wish 42

**Example Narrow Priority Objective.Wish** [Decision Criteria = **Decision Reasons**, Transparency Level = **{Data Trustworthiness, Software Engineering Process Transparency,Technical Inspection Transparency}**, Stakeholder Type = **Accident Investigator**, A/IS Incidents = **{Threatening Behaviour, Operator Failure}**, Incident Severity = **Lethal Force Activation**, A/IS Operation Types = **Cyber Defence**] @ 24 May 2022 : 42

*<- Tsg. Just an initial narrow parameters example.*

**Ambition Level:** transparent A/IS are ones in which it is possible to discover how and why a system made a particular decision, or in the case of a robot, acted the [illegible] 1

**Stakeholders:** A/IS TECHNOLOGISTS, CONSUMER ADVOCATES, ENGINEERS, ETHICISTS, GOVERNMENTS, IEEE P7001 Transparency Of Autonomo[illegible] Systems Standard, INDIVIDUALS, INTERNA...

**Scale:** % of [A/IS Operation Types] where the [Decision Criteria] has a [Transparency Level] for a [Stakeholder Type] for [A/IS Incidents] of [Incident Severity] 2

**Example Narrow Priority Current Level.Status: 24** [Stakeholder Type = **Accident Investigator**, A/IS Incidents = **{Threatening Behaviour, Operator Failure}**, Incident Severity = **Lethal Force Activation**, A/IS Operation Typ...

**Example Narrow Priority Objective.Wish: 42** [Decision Criteria = **Decision Reasons**, Transparency Level = **{Data Trustworthiness, Software Engineering Process Transparency,Technical Inspection Transparency}**, Sta...

**Relations:** IEEE STANDARD P7001 - Transparency Of Autonomous SystemS, Principle 1 — Human Rights.

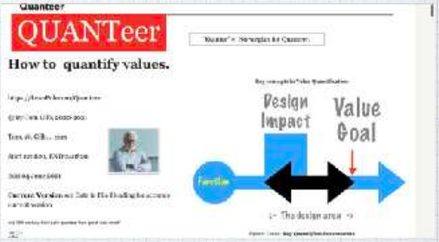
**Issue:** IssueActionIssue: How can we ensure that A/IS are transparent? 1

**IEEE2.Rationale:** Transparency is important to each stakeholder group for the following reasons:1. For users, transparency is important because it provides a simple way for them to understand what t...

**Candidate Recommendation IEEE2 .Implementation Plan:** Candidate RecommendationDevelop new standards* that describe measurable, testable levels of transparency, so that systems can be o...

**IEEE STD.Relations:** IEEE STANDARD P7001 - Transparency Of Autonomous SystemS, Principle 1 — Human Rights.

Main definition and Quantification




**Free copy.**

**https://tinyurl.com/Quanteer**

**Explains these quantification methods**

# 3.5 Multidimensional AI Models for Qualities (quantified), And Costs

Main definition and Quantification: details of dimensions

Candidate Recommendation

**Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined.**

Detail of the Transparency Scale of Measure, derived from the Ambition Level

**Ambition Level:** transparent A/IS are ones in which it is possible to discover how and why a system made a particular decision, or in the case of a robot, acted the way it did.

**Stakeholders:** A/IS TECHNOLOGISTS, CONSUMER ADVOCATES, ENGINEERS, ETHICISTS, GOVERNMENTS, IEEE P7001 Transparency Of Autonomous Systems Standard, INDIVIDUALS, INTERNATI..

**Tag.Scale:**

% of [A/IS Operation Types] where the [Decision Criteria] has a [Transparency Level] for a [Stakeholder Type] for [A/IS Incidents] of [Incident Severity]

Templates

**A/IS Incidents:** defined as:

Accidents, Misuse, Confusing Communication, Inexplicable Action, Threatening Behaviour, Financial Loss, Equipment Failure, Operator Failure, Data Corruption,

**A/IS Operation Types:** defined as:

Security Operation, Recovery Operation, Weapon Operation, Road Transport Operation, Flight Operation, Flight Monitoring. Operation, Decision-Making Operation, Cyber Attacking Operations, Cyber Attack, Cyber Defence,

**Decision Criteria:** defined as:

Right To Explanation, Decision Challenge, Decision Assessment, Decision Reasons,

**Incident Severity:** defined as:

Lethal Force Activation, Traceability Log Fault, Violations of Systems Embedded Community Norms, Mitigation Failure, Trust Violation,

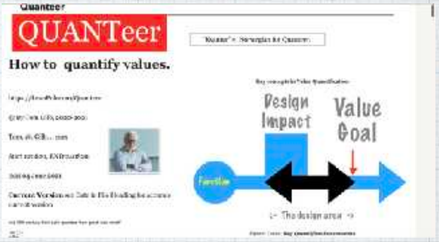
**Stakeholder Type:** defined as:

Accident Investigator, Validation Instance, Certification Instance, Private Person User, Responsible Component Provider, Responsible Systems Provider, Judges, Juries, Lawyers, Expert Witnesses, Court Process, Non Court Adjudication Process, Arbitration Process, Public Confidence,

**Transparency Level:** defined as:

Trusted User Level, Data Trustworthiness, Algorithmic Trustworthiness, Situational Awareness, Trace Records, Intentional Deception Justification, Systems Working Inspection, Potential Future Consequences of Action or Decision, Ordinary Language, Source Code, Transparency Respecting Security, Software Engineering Process Transparency, Technical Inspection Transparency,

Transparency Scale ←——Detailed Definition By Decomposition




Free copy. https://tinyurl.com/Quanteer

Explains these quantification methods

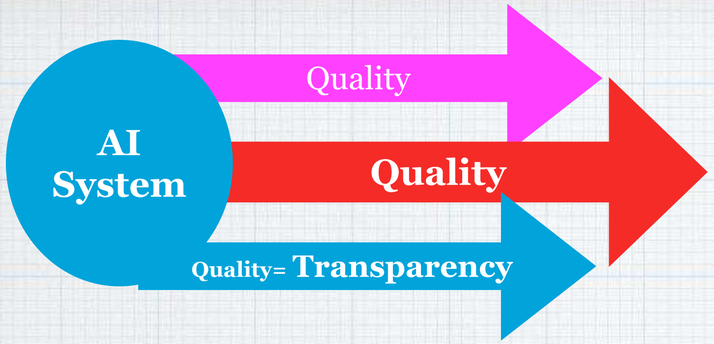
The example shows the use of [Scale Qualifiers] to define a quality at a high level
And to enable us to model very complex systems, with any combinations of [Scale Qualifiers]

**Scale:**
**% of [A/IS Operation Types]**

**where the [Decision Criteria]**

**has a [Transparency Level]**

**for a [Stakeholder Type]**
**for [A/IS Incidents]**
**of [Incident Severity].**




AI System

AI System Qualities

AI System Quality: Transparency

AI System Quality: Transparency, Scale

AI System Quality: Transparency, Scale
Qualifiers like [Transparency Level]

Qualifiers like [Transparency Level]

<- The Black Box

<- Our peep-holes into it

<- Specific peep-hole into it

<- Defining as a 'metric'

<- identifying a sub- dimension

<- Defining a Scale sub-dimension In terms of a set of Conditions

The result of this is that
Qualities are really well defined
We can model very complex AI systems
We can understand AI at a more-detailed level

Free copy. Explains details
https://tinyurl.com/Quanteer

# 3.6 Multidimensional AI Models for Qualities (quantified), And Costs

Candidate Recommendation

**IEEE**

**"Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined".**

Show Sidebar

**Principle 1 — Human Rights**

**Level:** Organization , **Status:** Not Determined **Type:** Value, **Labels:** no labels Edit

**Is Part Of:** GENERAL IEEE PRINCIPLES

Human Rights

Status 1 Wish 42

**Wish** [A/IS Influential Entities = **{Professional Bodies, Corporations,Startups}**, Consideration Type = **Real and Serious Ethics Technology Implementation**, A/IS System Types = **{Medical, Pharmaceutical,Educational,Media}**, Geographic Region = **All**] @ 30 Apr 2029 : **42** about a decade hence

*<- arbitrary but useful set of Scale Parameters*

**Ambition Level:** Human Rights need to be taken into consideration by various bodies who provide systems 1

**Stakeholders:** INDIVIDUALS, ORGANIZATIONS, PROFESSIONAL BODIES, RELIGIONS, RESEARCH, UN Declaration Of Human Rights .

**Scale:** % of [A/IS Influential Entities] who give a [Consideration Type] for [A/IS System Types] in a [Geographic Region]

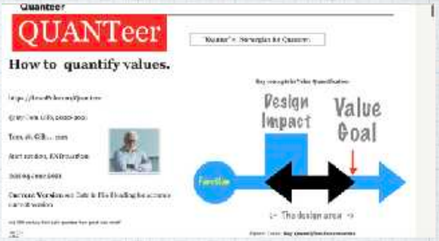
Main definition and Quantification

**Status: 1** [A/IS Influential Entities = **{Professional Bodies, Corporations,Startups}**, Consideration Type = **Real and Serious Ethics Technology Im** ... **dical, Pharmaceutic**...

**Wish: 42** about a decade hence [A/IS Influential Entities = **{Professional Bodies, Corporations,Startups}**, Consideration Type = **Real and Serious Ethics Technology Implementation**, A/IS System Types ...

**Issue:** IssueActionIssue:How can we ensure that A/IS do not infringe upon human rights? 1

**Relations:** Principle 4 — Transparency.




**Free copy.**
**https://tinyurl.com/Quanteer**

**Explains these quantification methods**

The App used here is https://www.gilb.com/valplan

# 3.7 Multidimensional AI Models for Qualities (quantified), And Costs

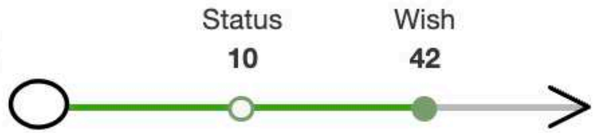
**Principle 2 — Prioritizing Well-Being**

'Well Being' 1-page Objective

**Level:** Organization , **Status:** Not Determined **Type:** Value, **Labels:** no labels Edit

Show Sidebar

**Is Part Of:** GENERAL IEEE PRINCIPLES




**Wish** [Human Type = **{Child. Teen, Female,Strong,Poor}**, Satisfaction Type = **Largely Satisfied**, Life Type = **Student**, Conditions Of Life = **{Health, Education}**, Balance Effect Ratio = **Positive Effects Dominate**] @ 09 Apr 2029 : **42** *<- Tom Gilb. Random selection of Scale Parameters to illustrate the use of them to define levels, and dimensions.*

**Ambition Level:** encompassing human satisfaction with life and the conditions of life as well as an appropriate balance between positive and negative affect. 1

**Stakeholders:**

**Scale:** % [Human Type] [Satisfaction Type] with [Life Type] and [Conditions Of Life] and a [Balance Effect Ratio] 1

Main definition and Quantification

**Status: 10** [Human Type = **{Child. Teen, Female,Strong,Poor}**, Satisfaction Type = **Largely Satisfied**, Life Type = **Student**, Conditions Of Life = **{Health, Education}**, Balance Effect Ratio = **Positive Effects Domina…**

**Wish: 42** [Human Type = **{Child. Teen, Female,Strong,Poor}**, Satisfaction Type = **Largely Satisfied**, Life Type = **Student**, Conditions Of Life = **{Health, Education}**, Balance Effect Ratio = **Positive Effects Dominate.**

**Note:** This entire arbitrary construction, of the Well-Being specification is not intended as a correct or complete, or agree definition. On the contrary, it is intended to show1. paths we mig…

**Free copy.**
**https://tinyurl.com/Quanteer**
**Explains these quantification methods**

# 3.8 Multidimensional AI Models for Qualities (quantified), And Costs

Candidate Recommendation

**Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined.**

## DETAIL OF Scale for Prioritising Well Being

**Ambition Level:** encompassing human satisfaction with life and the conditions of life as well as an appropriate balance between positive

**Stakeholders:**

**Tag.Scale:**

% [Human Type] [Satisfaction Type] with [Life Type] and [Conditions Of Life] and a [Balance Effect Ratio]

Templates ⌘+e for

**Balance Effect Ratio:** defined as:

Positive Effects Dominate, About Equal Positive and Negative Effects, Negative Effects Dominate

**Conditions Of Life:** defined as:

Health, Family, Wealth, Security, Education, Mobility, Freedom of Religion, Freedom of Opinion, Freedom of Association, Equal Opportunity, ...

**Human Type:** defined as:

Child. Teen, Adult, Retired, Male, Female, Other Sex, Strong, Weak, Rich, Poor, Handicapped

**Life Type:** defined as:

Student, Working, Retired, Volunteer, In Care

**Satisfaction Type:** defined as:

Extremely Satisfied, Largely Satisfied, Dissatisfied, Extremely Dissatisfied

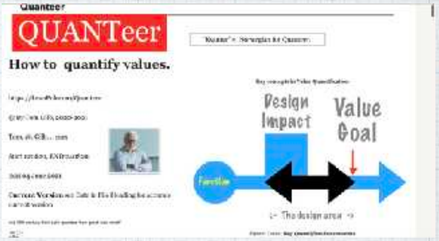
**Target Time Units:**

Calendar Date

Advanced...

**Source:** by tomgilb - Apr 30th 2019, 00:26

## 'Well Being' Scale of measure definition




**Free copy.**
**https://tinyurl.com/Quanteer**

**Explains these quantification methods**

# 3.9 Multidimensional AI Models for Qualities (quantified), And Costs

**Principle 3 — Accountability**

**Level:** Organization , **Status:** Not Determined **Type:** Value, **Labels:** no labels Edit

**Is Part Of:** GENERAL IEEE PRINCIPLES


**Ambition Level:** Based on the cultural context, application, and use of A/IS, people and institutions need clarity around the manufacture and deployment of these systems to establish res...

**Issue:** IssueAction->How can we assure that designers, manufacturers, owners, and operators of A/IS are responsible and accountableMaybe see Candidate Recommendations IEEE p ...

**Stakeholders:** CULTURE, INDIVIDUALS, ORGANIZATIONS.

**Scale:** % of cases of [Harm] where we can establish [Responsibility] towards [People And Institutions] for a given [A/IS Application]

**Status: 0** When 29 Apr 2019

**Wish: 0** When ?

Accountability 1-page Objective

Main definition and Quantification

Free copy.
https://tinyurl.com/Quanteer
Explains these quantification methods

# 3.10 ‘ACCOUNTABILITY Multidimensional AI Models for Qualities (quantified), And Costs

**Ambition Level:** Based on the cultural context, application, and use of A/IS, people and institutions need clarity around the manufacture and deployment of these systems to

**Issue:** IssueAction->How can we assure that designers, manufacturers, owners, and operators of A/IS are responsible and accountableMaybe see Candidate Recommendat

**Stakeholders:** CULTURE, INDIVIDUALS, ORGANIZATIONS.

***Tag*.Scale:**

% of cases of [Harm] where we can establish [Responsibility] towards [People And Institutions] for a given [A/IS Application]

Main definition and Quantification: details dimensions

Templates

for formatting, to add qualifiers / terms

Free copy:
https://tinyurl.com/Quanteer
Explains these quantification methods
IF YOU CANNOT LEARN THIS,
MOVE ASIDE, ChatGPT is
Very good at it. TG 2024

Accountability Scale detail

**A/IS Application:** defined as:

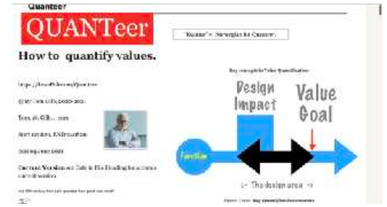
Transport, Advice, Medical, Education, Household, Care Institutions, ...

**Harm:** defined as:

Potential Harm, Actual Harm, Severe Harm, Mass Harm, Fatal Harm, Psychological Harm

**People And Institutions:** defined as:

Individual Stakeholders, Institutional Stakeholders, Innocent Victims, ...




**Responsibility:** defined as:

Legal Culpability, Ethical Responsibility, Potential Responsibility, Acknowledged Responsibility, Indirect Responsibility, Financial Responsibility, Mitigation Responsibility

**Target Time Units:**

Calendar Date

Advanced...

# Summary thus far

All AI-system qualities,
and other stakeholder values

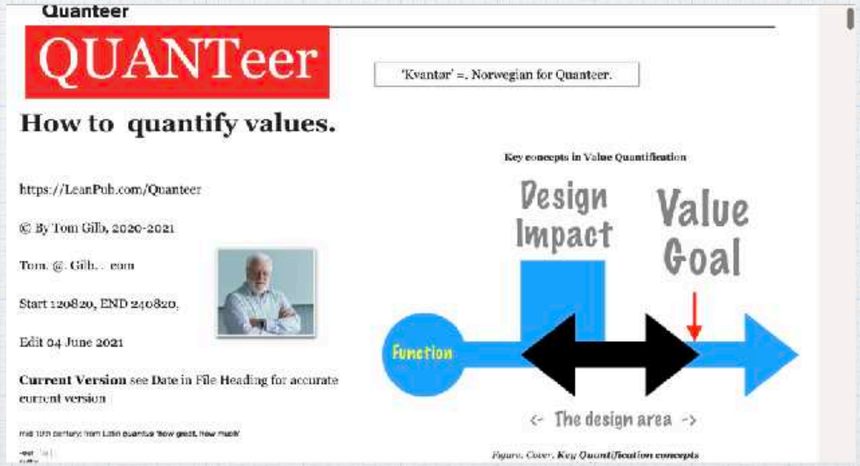
can be defined as - a quantity, a 'metric'.

So that we have a clearer, stable, tool
for comparing - AI-system performance.

**Here is a free detailed process for doing that quantification**

**Without exception**

**https://tinyurl.com/Quanteer**



# 4.0 The AI Stakeholder Model

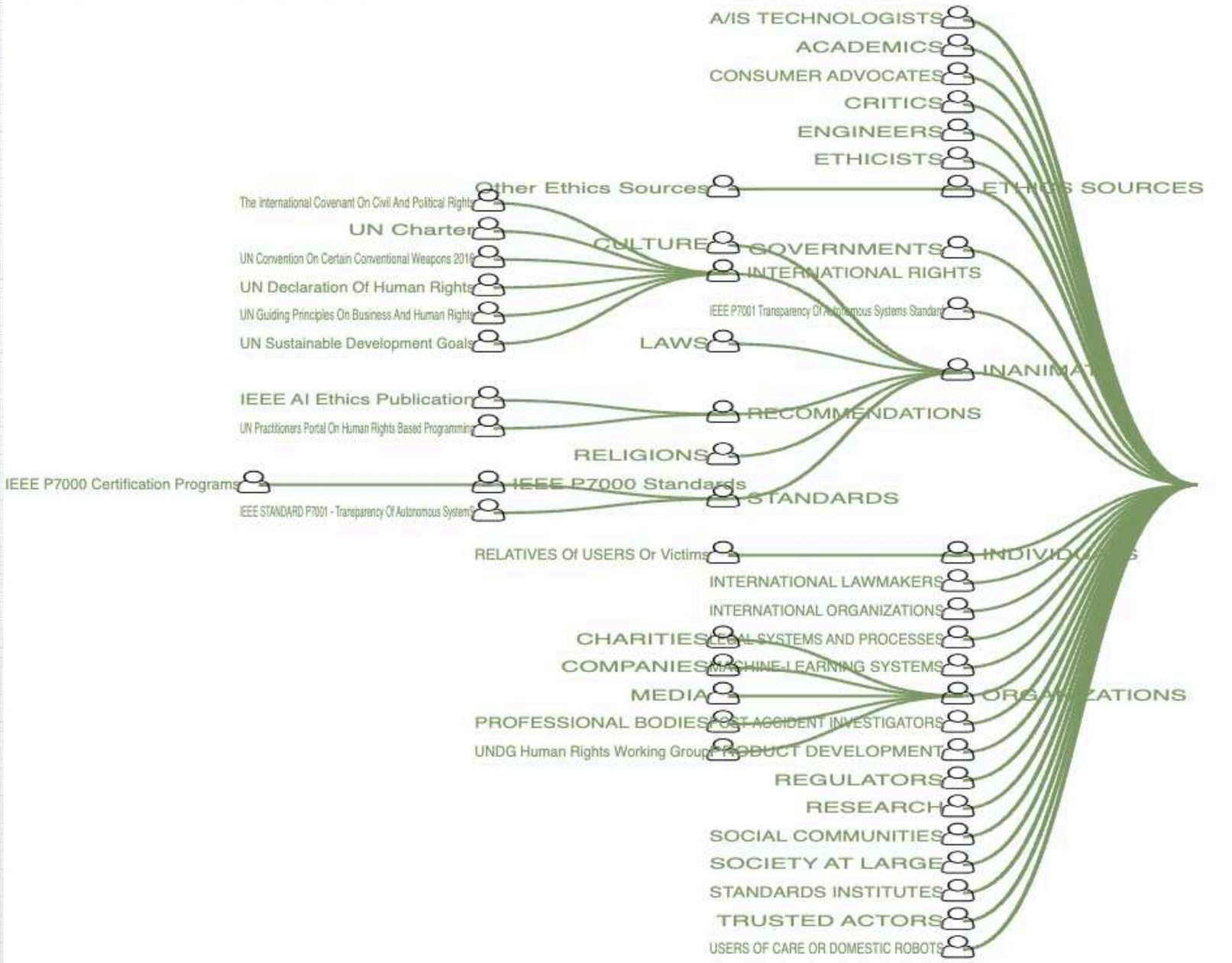
* Another '**Technoscope**'** for understanding an AI-system, is by analyzing, and listing, all critical, or interesting, stakeholders.
* Stakeholders are the **direct sources of requirements** for the AI-system
* against which, various AI-systems can be **evaluated**
* **Is an AI-system the best for your entire set of critical stakeholders?**


Stakeholder Engineering.
By Tom Gilb
Leanpub.com/StakeholderEngineering
Released 27 July 2021, Leanpub, 177 pages.
#stakeholders #systemsengineering #requirementsengineering
Gilb, T. Free copy
https://tinyurl.com/StakeholderBook

** FREE COPY BOOK : TECHNOSCOPES. https://www.gilb.com/offers/YYAMFQBH/

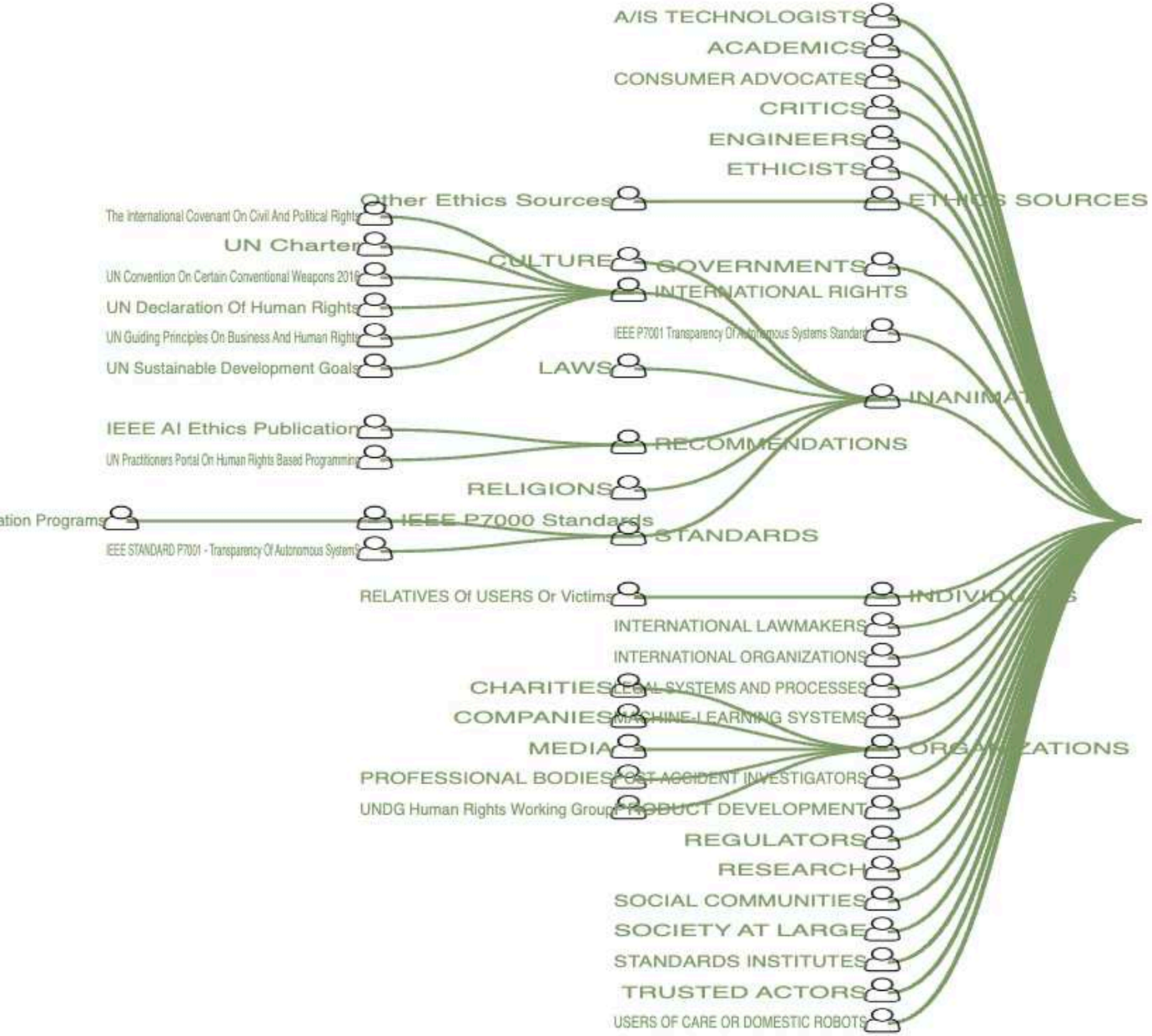
# 4.1 The AI Stakeholder Model

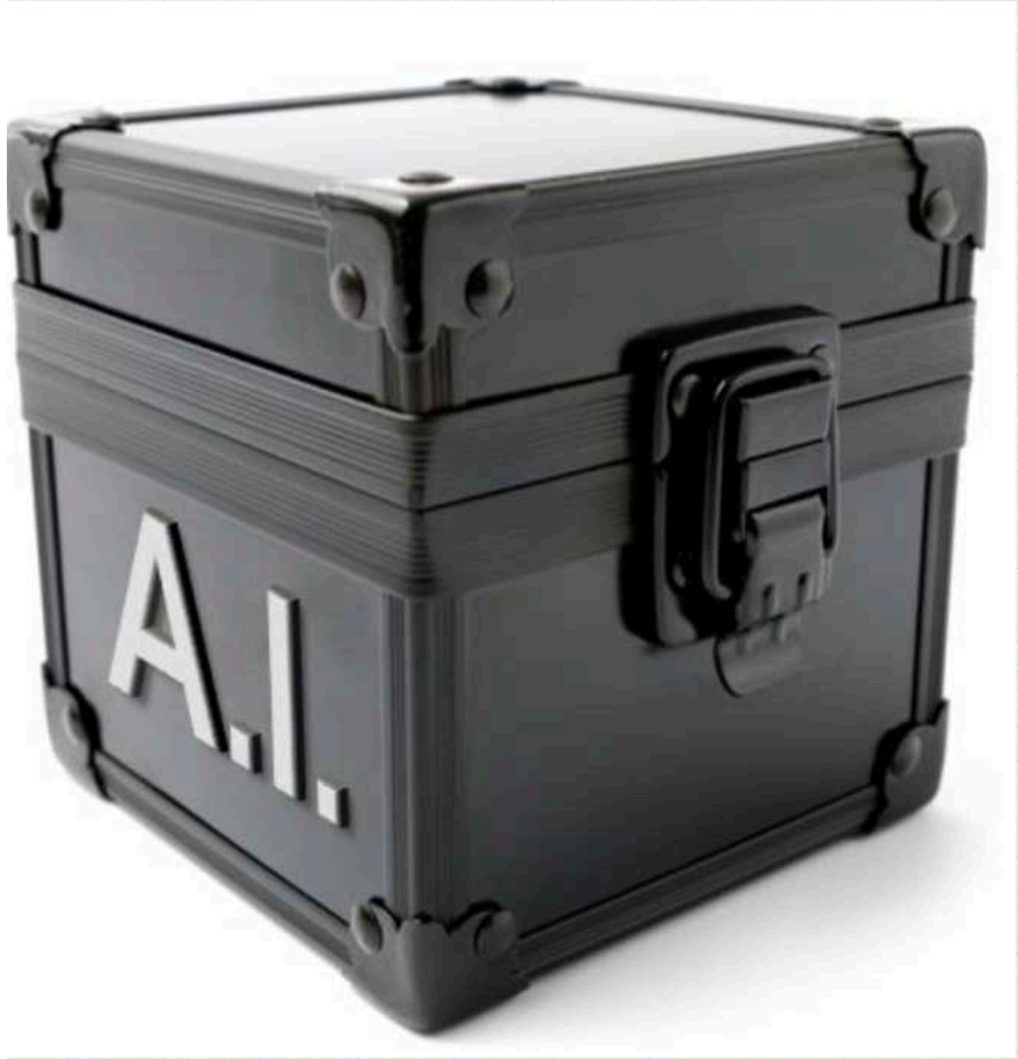
- Another 'Technoscope' for understanding an AI-system is by analyzing and listing all critical or interesting stakeholders.
- These are the direct sources of requirements for the AI-system
- Against which various AI-systems can be evaluated




IEEE AI Ethics Standard, Stakeholder Analysis (in ValPlan app)

# 4.2 The AI Stakeholder Model (a tool for understanding the AI Black Box)




IEEE AI-Ethics Standard Stakeholder Analysis (in ValPlan app)

# 4.3 The AI Stakeholder Model

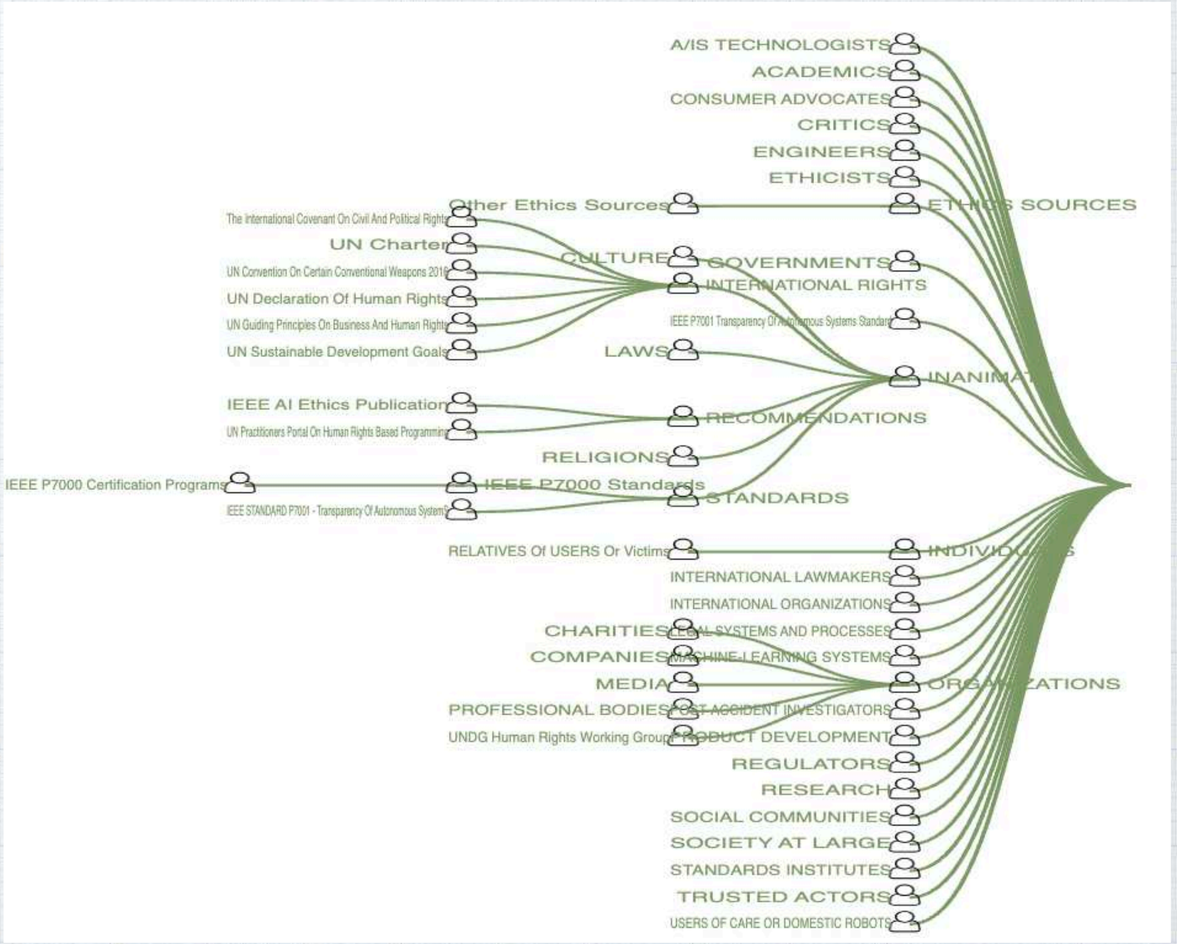
## Why does a stakeholder model help us to understand the AI-system?

1. It helps you **discover stakeholder needs**, and specific stakeholder levels, and types (who, what, where) of those needs (not just a generic one-size-fits all needs)

2. It helps **evaluate the AI system from the point of view of the specific stakeholders**

3. It gives you a chance to start an **early** evolutionary incremental **value flow, towards those stakeholders.**

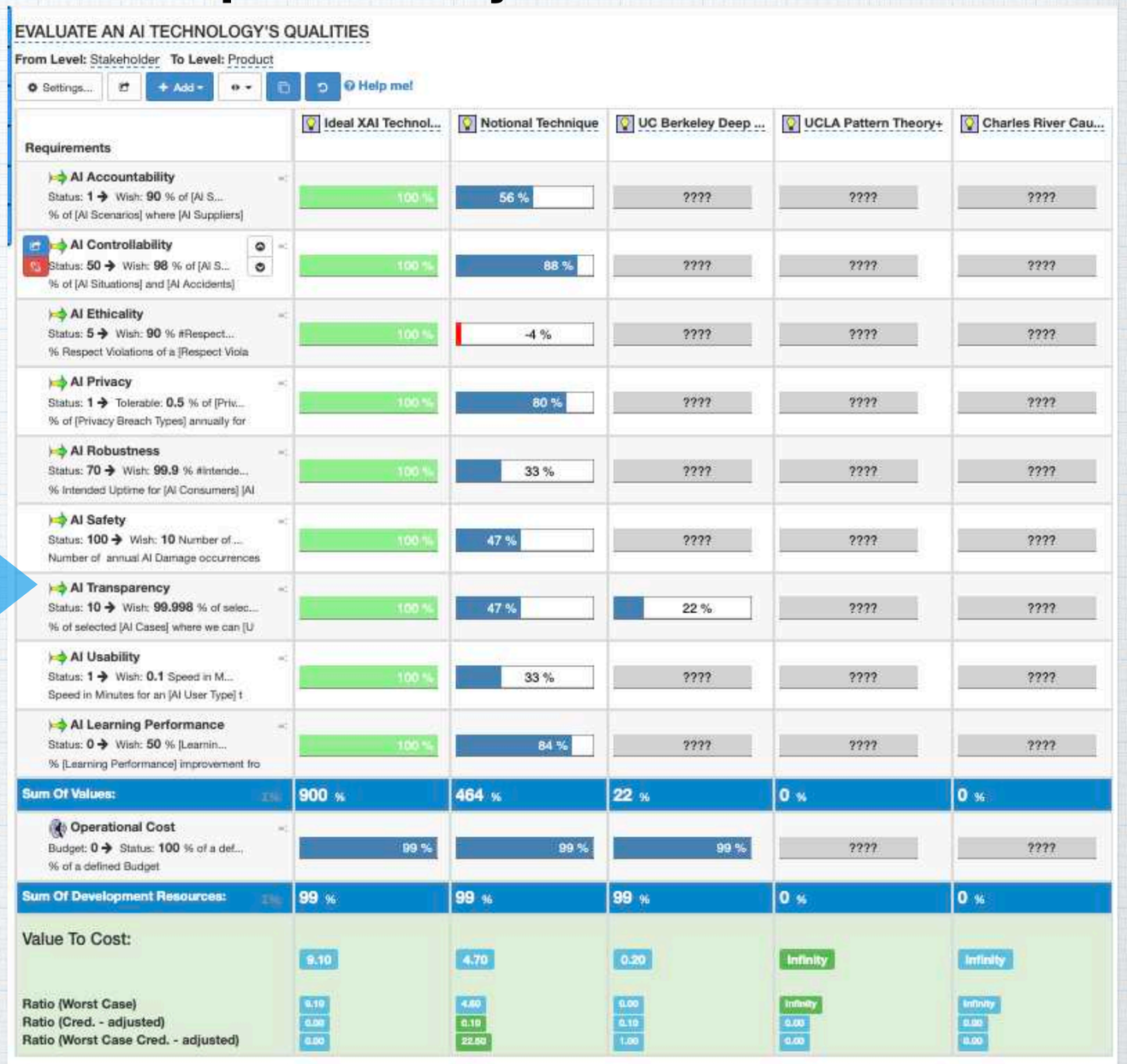
4. It helps **identify 'hard constraints'**, like laws and budgets, which would make selection of a particular AI-system invalid or a big mistake.




IEEE Standard Stakeholder Analysis (in ValPlan app)

# 5. Understanding AI Strategies in light of Multiple Quality and Cost Attributes

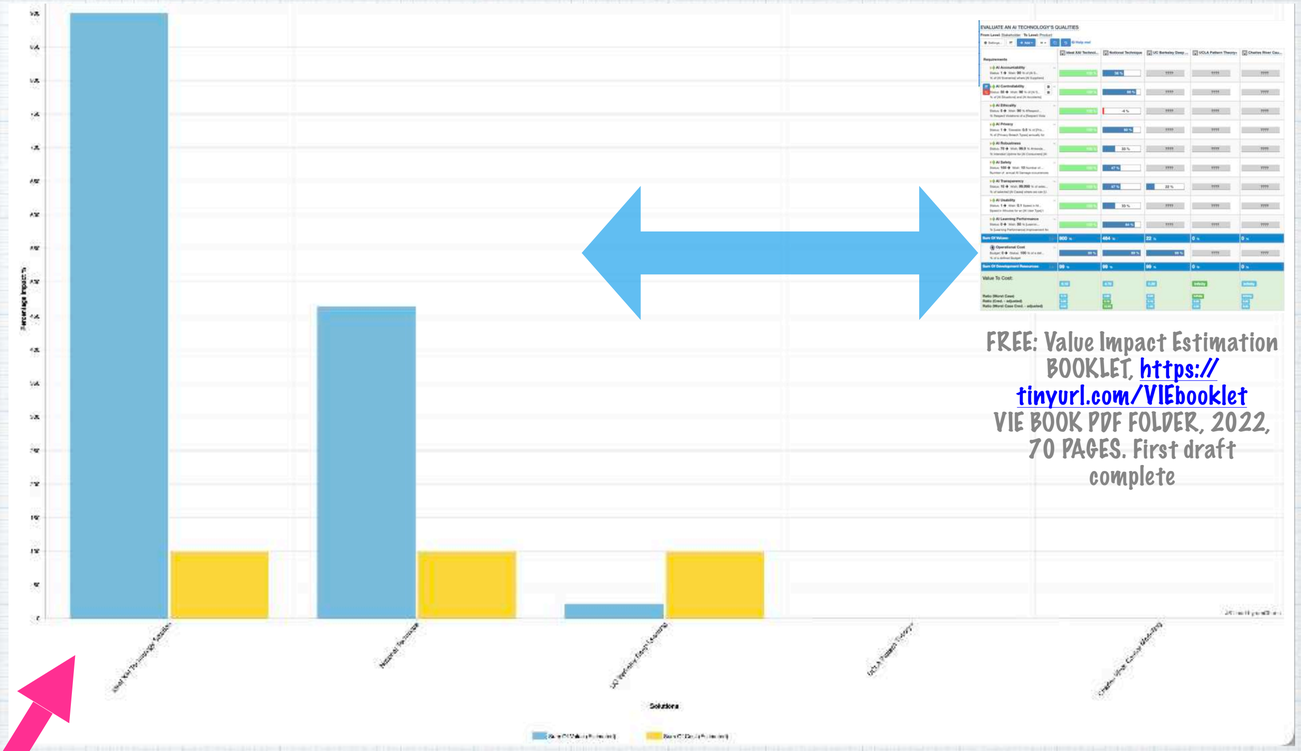
- We can estimate the effectiveness and costs of AI-systems against all value objectives and costs
- Because, if we quantify our AI Stakeholder Values,
  - we have some clear idea of what we need to choose; in terms of an AI-system




FREE: Value Impact Estimation BOOKLET, https://tinyurl.com/VIEbooklet
VIE BOOK PDF FOLDER, 2022, 70 PAGES. First draft complete

# 5.1 Understanding AI Strategies in light of Multiple Quality and Cost Attributes

- We can estimate the effectiveness and costs of AI-systems against all value objectives and costs
- Because, if we quantify our AI Stakeholder Values,
  - we have some clear idea of what we need to choose; in terms of an AI-system



FREE: Value Impact Estimation BOOKLET, https://tinyurl.com/VIEbooklet
VIE BOOK PDF FOLDER, 2022, 70 PAGES. First draft complete

The sum of AI-system Values and (Yellow) Sum of System Costs

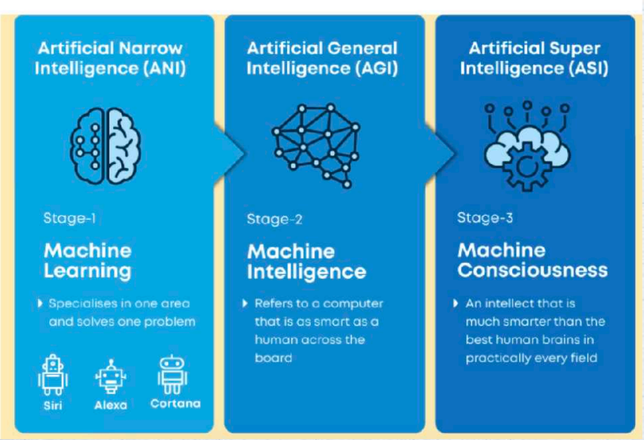
Simplifying the presentation of the attributes of an AI-system

# 6. Teaching The Multi-AI model to students and research students as a tools for researching AI developments. AN ‘ACADEMIC’ VIEWPOINT


* Teach ‘Planguage’ as a basic tool for ‘life’, and all AI research.
  * See https://www.researchgate.net/publication/237129623_Competitive_Engineering_A_Handbook_for_Systems_Engineering
    * and get free copy of ‘Competitive Engineering’
* Use Planguage to analyze AI systems and write papers and thesis
* To Be Clear: I am not here to sell books or services, I am retired, and I want to convince you to try my ideas, and I want to give you free written materials to teach them.

# 7. The Quality distinctions between Large Language Model (LLM) AI and the next AI generation (AGI, Artificial General Intelligence)

- Large Language Models: merely process data and spit out tailored versions of it, like 'magic'. Often wrong and fake news.
- Artificial General Intelligence: is Not based on Big Data.
  - They are based on teaching AI to 'reason like a human' and solve interesting problems. The AI future is AGI
    - See graphmetrix.com for emergence of a real product that already does this on building drawings and papers. (Disclaimer: I am investor here)
- There are some videos on site to explain
- Peter Lawrence explains this (AGI technicalities):
  - https://medium.com/@peter.lawrence_47665/knowledge-graphs-large-language-models-the-ability-for-users-to-ask-their-own-questions-e4afc348fa72




https://www.mygreatlearning.com/

# AN AI-PRODUCER OPINION

**My one word response to this ("The AI Dilemma") is "delusional".**

**The fact is, there is no Intelligence at all in LLMs,**

**and because of this, they can't do the intelligent things people think they can**

**because they use words that were output from intelligence.**

**There is A for sure.**

**The primary danger of these things is fooling people,**
**so a big productivity boost for criminality.**

**Outside of that, LLMs are literally pretty useless for anything mission critical that needs to work autonomously.**

**So these guys are crazy in thinking that anyone will lose their jobs, the whole economy will change etc. For sure the criminals will be able to commit more crime more easily however.**

**Those radiologists that were supposed to be out of work by now... those truck drivers, taxi drivers... etc etc. Its crazy hype where billions will be lost as has happened with autonomous driving.**

**Intelligence is needed for these to work, and that is fully missing.**

**(Except for what we are working on ... ;) <-GRAPHMETRIX.COM**

Fred Gibson
*Founder & CEO*
*mobile: 415.335.8232*

*APRIL 11 2023*

https://rocketreach.co/frederick-gibson-email_7244557

https://vimeo.com/809258916/92b420d98a The AI Dilemma

# AN AI-PRODUCER OPINION

Email 14 April 2023

I would say that the most important point is
that LLMs and ml/dl have **no intelligence at all,**

and naming them "Artificial Intelligence" is both **false advertising** and promising something the systems can never deliver, intelligence.

They can get you to mission-critical (and very useful) perception *if* the input is pointed at the **world of space-time,**

that is measurements that an **intelligent system can then use to reason with,**

but pointing the input at words **bypasses this perception potential** entirely
and gives you just **word pattern outputs of random usefulness.**

Fred Gibson
Founder & CEO
mobile: 415.335.8232

ml=Machine Learning, dl= Deep Learning, LLM= Large Language Model

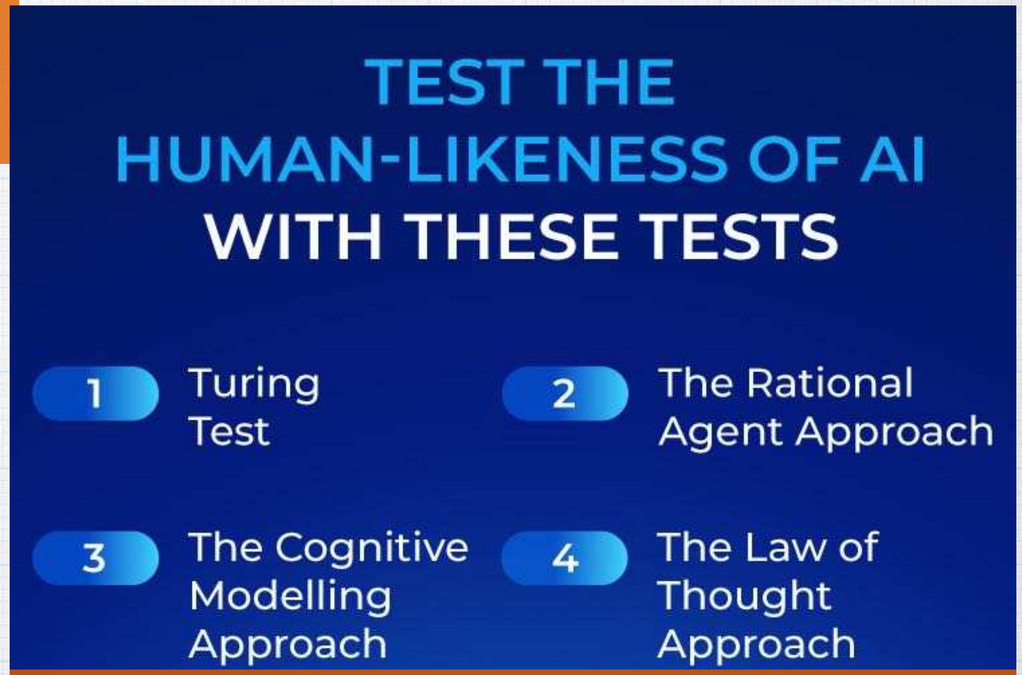
https://rocketreach.co/frederick-gibson-email_7244557

# How do we measure to see if Artificial Intelligence is acting like a human?

Even if we reach that state where an AI can behave as a human does, how can we be sure it can continue to behave that way?

We can base the human-likeness of an AI entity on the:

- Turing Test
- The Cognitive Modelling Approach
- The Law of Thought Approach
- The Rational Agent Approach




https://www.mygreatlearning.com/blog/what-is-artificial-intelligence/

# 8. The Penta Model as a basic high level view of any AI system.

* The Penta Model is a high-level simplification of the analysis ideas in Planguage (Competitive Engineering, gilb.com), https://tinyurl.com/PentaPaper

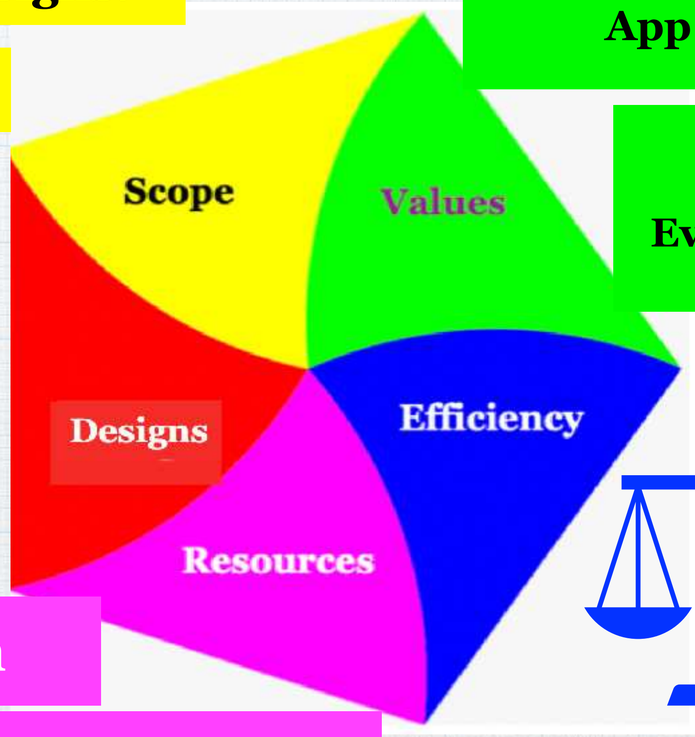
**Funct Tax: Compute all relevant Taxes**

**Design Constraint A: Avoid Patented Designs.**

**Function Constraint P: Avoid Political Bias**

Strategy 1: No Cure No Pay Contract

Design D: Digital Twin




**Quality Usability:**
**App user can sign up in avg. 5 minutes.**

**AI Decision Traceability:**
**Everyone can access Decision Reasoning sa**

**AI Decision Traceability &**

**Quality Usability**

over /

**Deadline & Budget**

**Budget: € 1 million**

**Deadline: : January 1 Next Year**

# 9. The Principles of AI Understanding, outside the blackbox.



1. We cannot hope to translate AI detailed complexity into understanding of qualities and costs

2. We have to work outside the AI-Black-Box with multiple quantified qualities and costs.

3. We can compare AI-systems, against their known ability, to satisfy quality objectives and cost constraints, by using an Impact Estimation Table. https://tinyurl.com/VIEbooklet. And we can use AI-LLMs to draft stakeholders, values, & costs, tables.

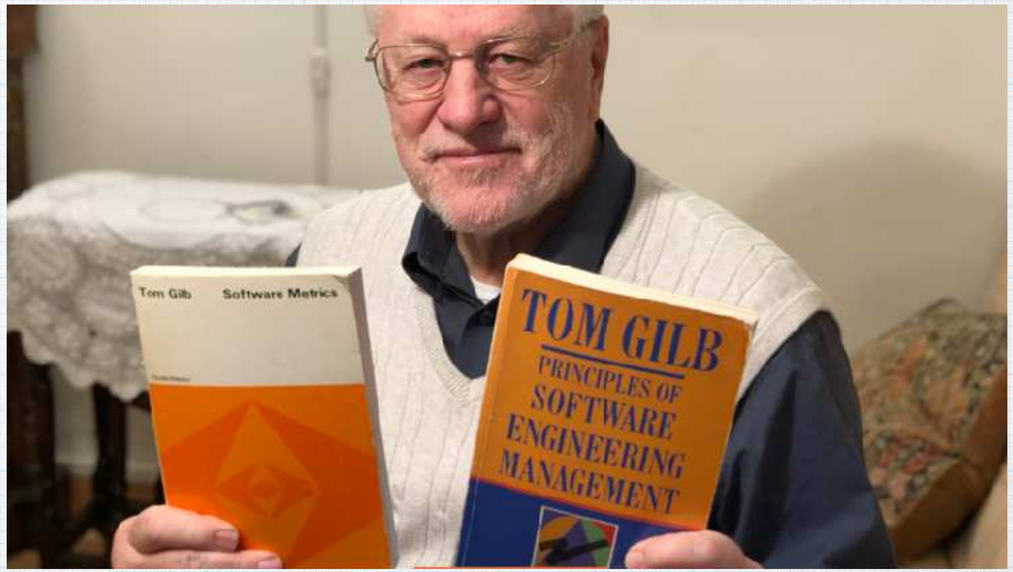
4. Identifying critical AI stakeholders gives better understanding of AI requirements, and gives opportunities for early results delivery, to selected segments.

5. If you cannot get real evidence about AI-system qualities and costs, then you have got specific 'known unknowns', risks. Which (the known unknowns) might be made into more knowledge, by experimentation, testing, and pilot use.

# LAST SLIDE

- tom@gilb.com
- Www.gilb.com
- www.graphmetrix.com (Investor, Advisory Board)
- @ImTomGilb
- http://www.linkedin.com/in/tomgilb
- twitter.com/imtomgilb
- Location: Near Oslo Norway
- Retired (83), but enjoy discussing and sharing
- Honorary Fellow of BCS (2013)
- https://www.researchgate.net/profile/Tom-Gilb
- https://leanpub.com/u/tomgilb

SLIDES FOLDER




I published 'Software Metrics' in 1976 - and been fighting for a metrics (engineering) culture, ever since.

AI, Security, and IT need metrics badly.

I am only 83 (July 2024), so I will irritate 'poetic methods' people for quite a while

And after that my writings will fight on!

You probably do not have time in a 45 minute lecture to go deeper
BUT
Here is more technical detail on quantifying AI qualities

https://www.researchgate.net/publication/381932584_eXplaining_AI_A_Serious_'Multi-dimensional_Metrics_Attack'_on_Poor_AI_'Academic_and_Standards'_Thinking_Planning

XAI Talk 2019 Slides


# XAI SLIDES

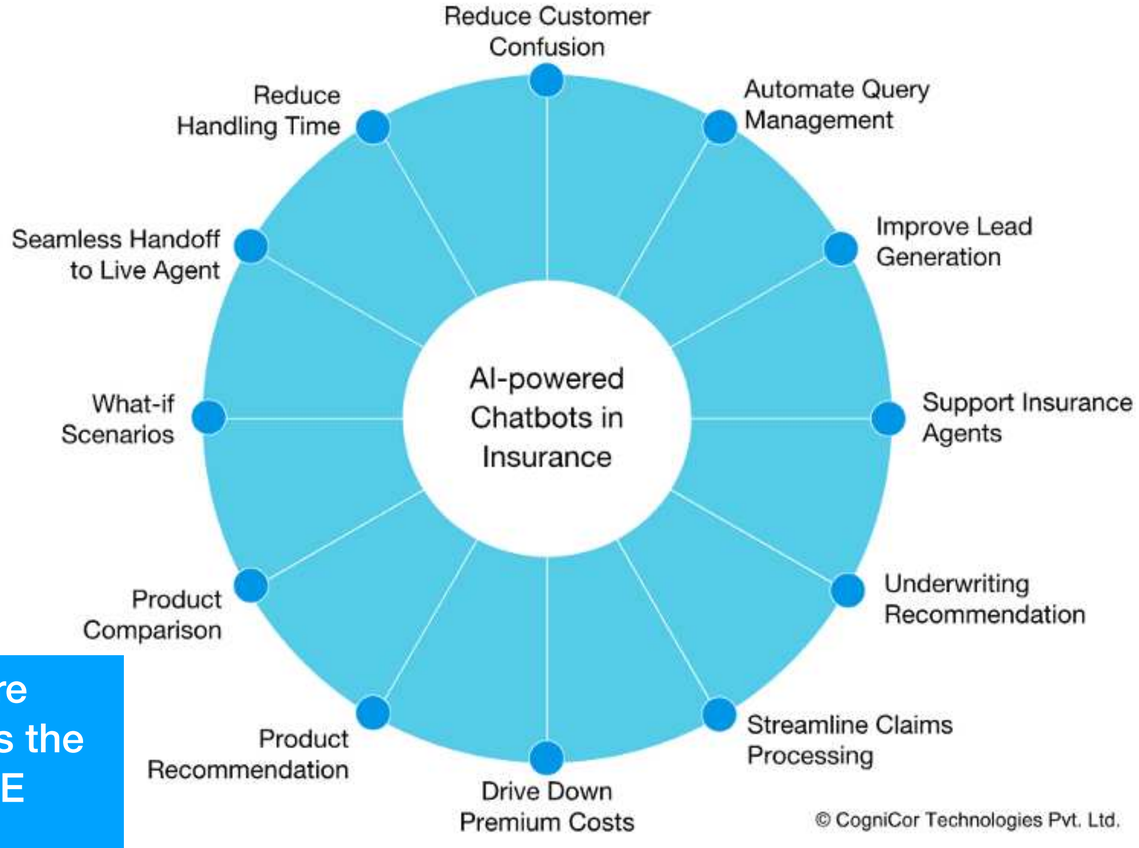
XAI Talk 2019 Slides

These are detailed quantifications of the AI Qualities in the White House presentation 2016
See slides
https://www.researchgate.net/publication/381932584_eXplaining_AI_A_Serious_'Multi-dimensional_Metrics_Attack'_on_Poor_AI_'Academic_and_Standards'_Thinking_Planning

# ‘XAI’

# “eXplaining A.I.”

## A Serious ‘Multi-dimensional Metrics Attack’ on Poor AI ‘Academic and Standards’ Thinking & Planning




I like this figure because it shows the MEASURABLE POTENTIAL AI Values IN A PARTICULAR DOMAIN (Insurance)


Tom Gilb

tom@Gilb.com


For Gilbfest London 25th June 2019, at 15:50- 16:35

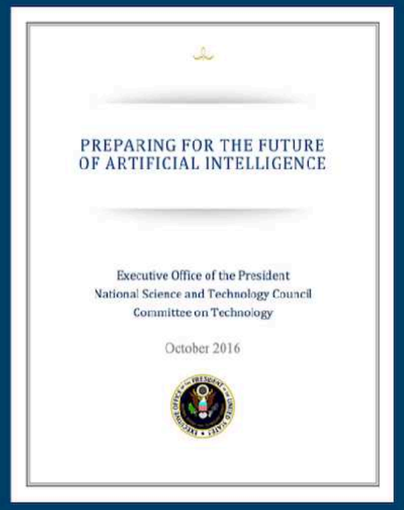
https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

**‘Principles’ = ‘Qualities’ !**

**With the *usual lack of quantified definition***

# Oct. 2016 : “Preparing for the future of AI” from USA

IBM




Distributed material

## Proposal of Discussion toward Formulation of AI R&D Guideline

Referring OECD guidelines governing privacy, security, and so on, **it is necessary to begin discussions and considerations toward formulating an international guideline consisting of principles governing R&D of AI to be networked** (“**AI R&D Guideline**”) as framework taken into account of in R&D of AI to be networked.

**Proposed Principles in “AI R&D Guideline”**

1. **Principle of Transparency**
   Ensuring the abilities to explain and verify the behaviors of the AI network system
2. **Principle of User Assistance**
   Giving consideration so that the AI network system can assist users and appropriately provide users with opportunities to make choices
3. **Principle of Controllability**
   Ensuring controllability of the AI network system by humans
4. **Principle of Security**
   Ensuring the robustness and dependability of the AI network system
5. **Principle of Safety**
   Giving consideration so that the AI network system will not cause danger to the lives/bodies of users and third parties
6. **Principle of Privacy**
   Giving consideration so that the AI network system will not infringe the privacy of users and third parties
7. **Principle of Ethics**
   Respecting human dignity and individuals’ autonomy in conducting research and development of AI to be networked
8. **Principle of Accountability**
   Accomplishing accountability to related stakeholders such as users by researchers/developers of AI to be networked

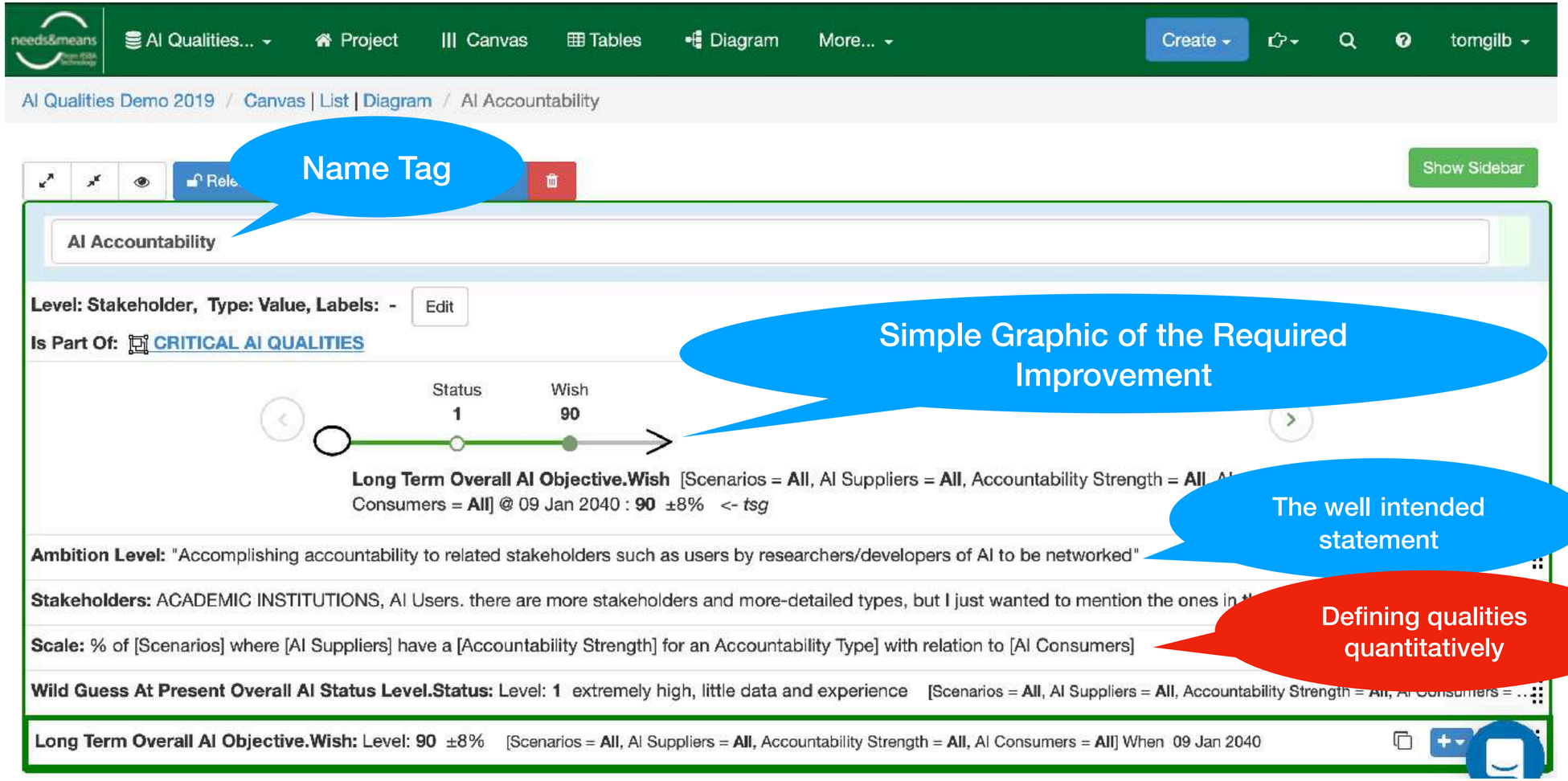


17

# AI Accountability


**https://www.smartcompany.com.au/people-human-resources/the-a-list-six-steps-to-accountability-in-your-business/**

# AI Accountability Defined
# Fuzziness -> Quantified. Structured, Enriched

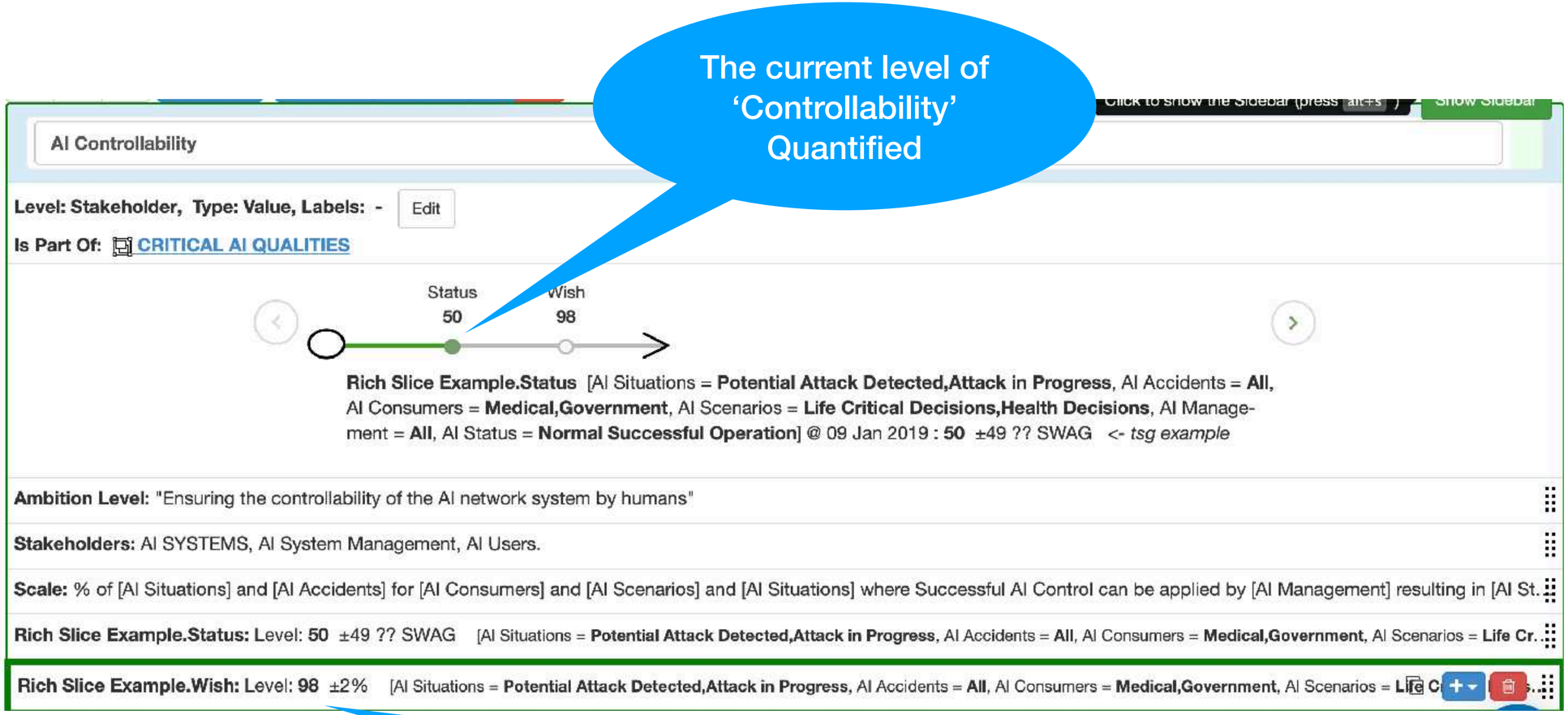


# AI Accountability Scale
# Detailed Scale Parameters
# Why is this Useful?

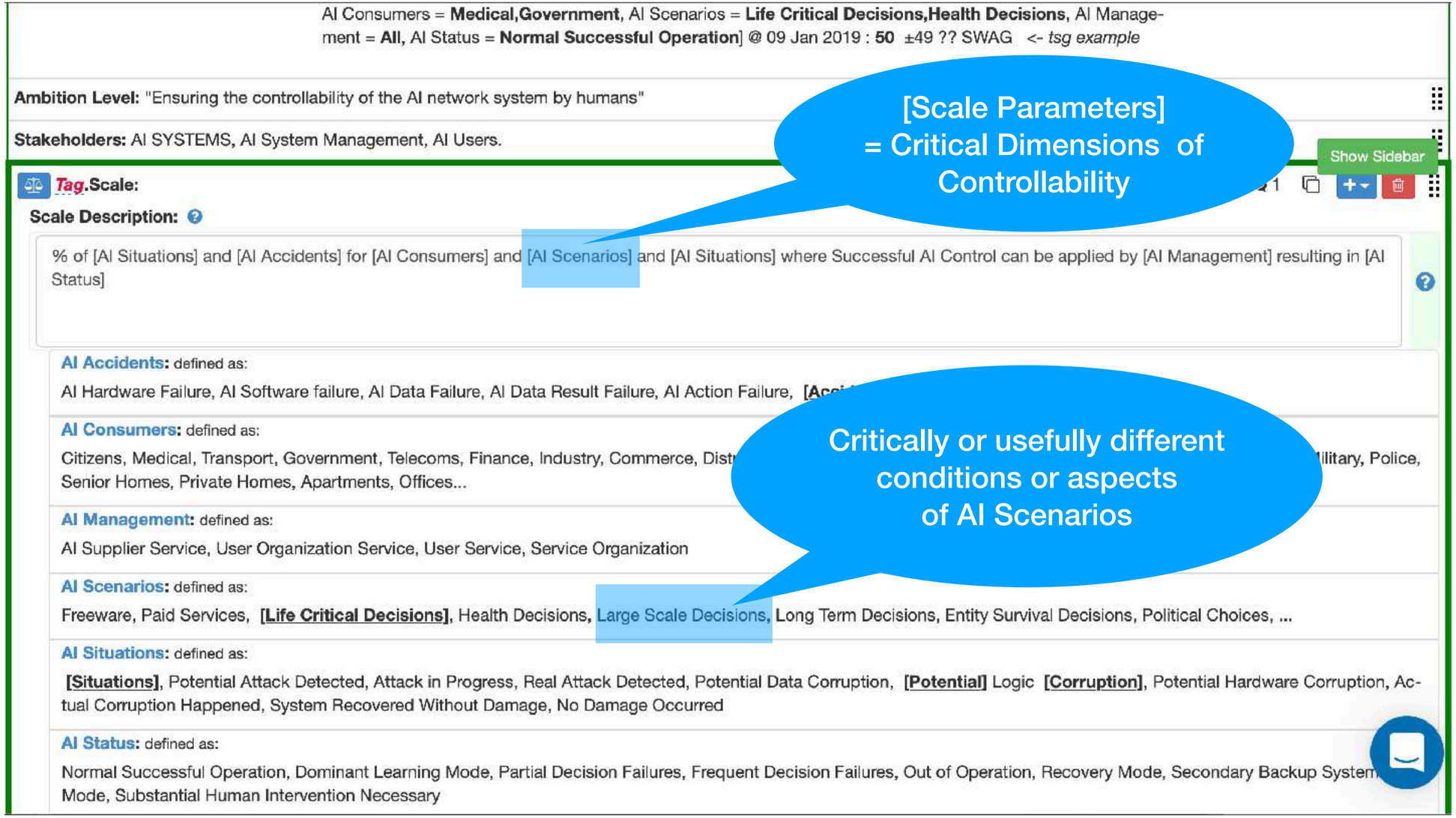

Slide error note: lack of [Accountability Type]

# AI Controllability

# AI Controllability: Quantified



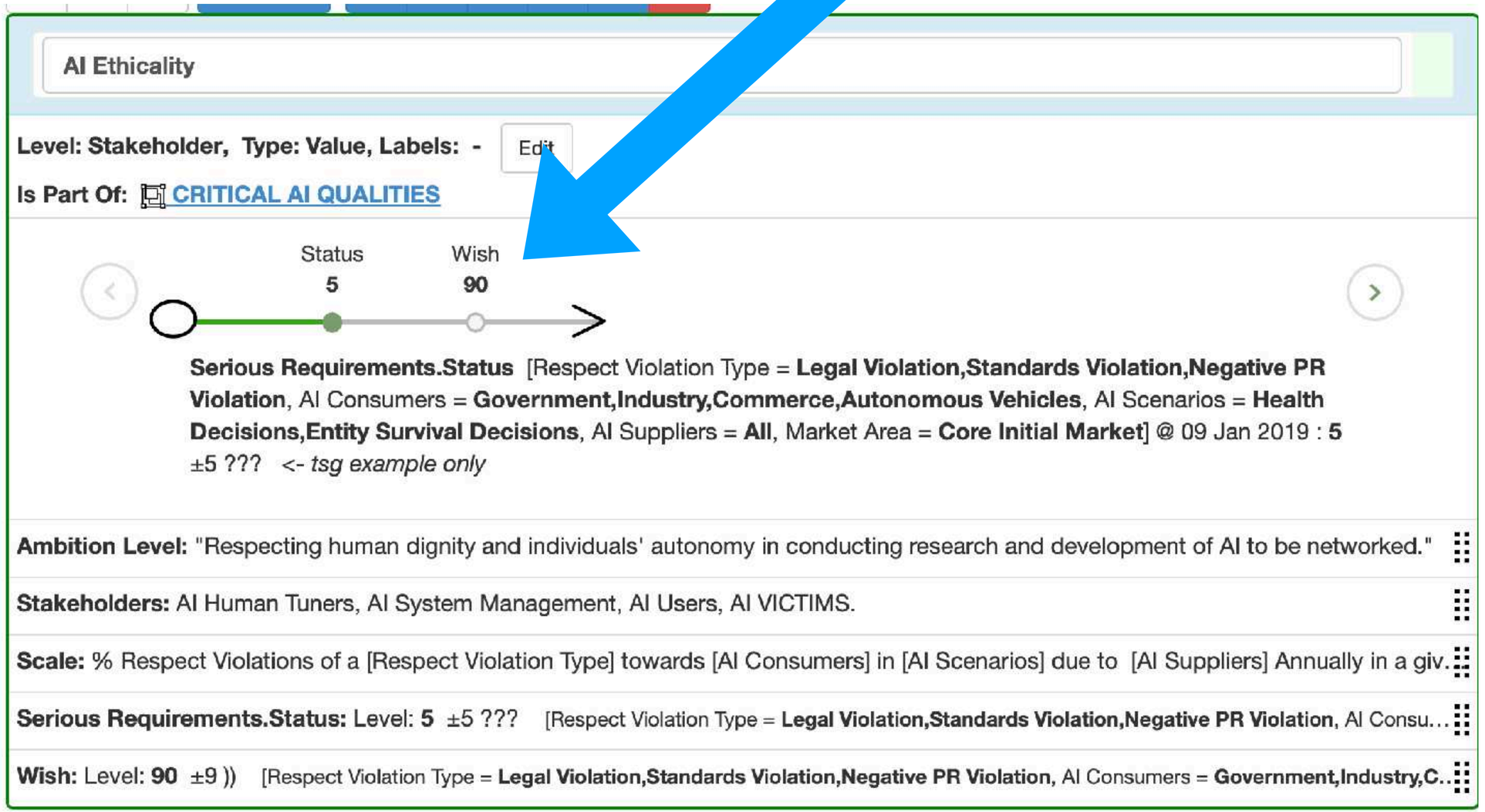
# AI Controllability: Scale of Measure Detail



# AI Ethicality

https://medium.com/salesforce-ux/how-to-build-ethics-into-ai-part-i-bf35494cce9

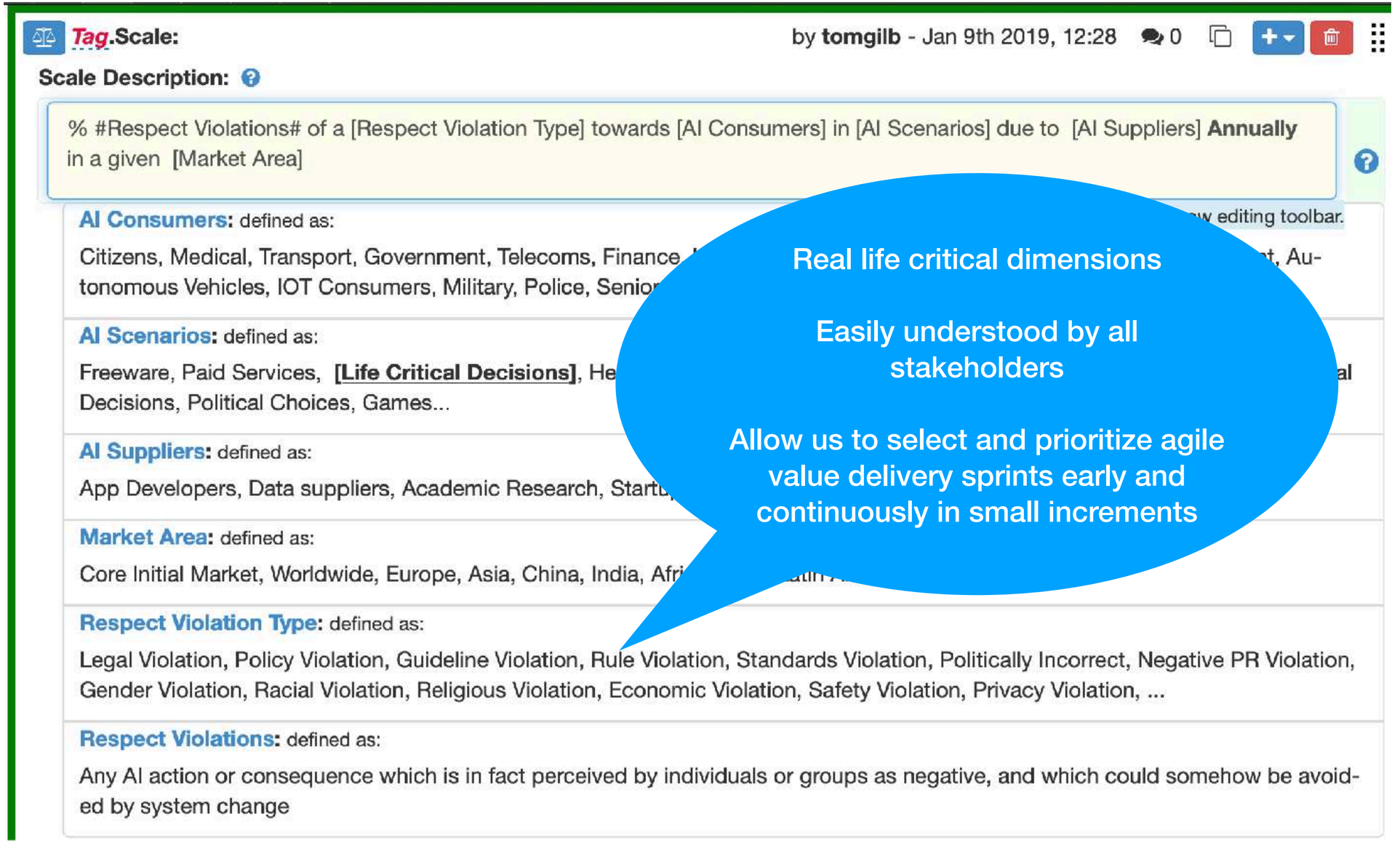
# AI Ethicality: Quantfied



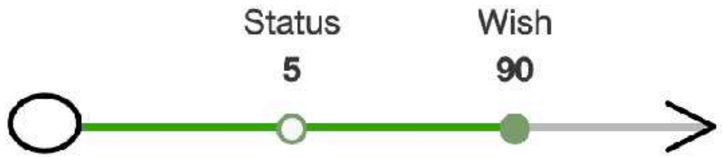
# AI Ethicality:
# Summary and Ethicality Scale of Measure

6 Dimensions

# AI Ethicality: Scale Parameter Detail



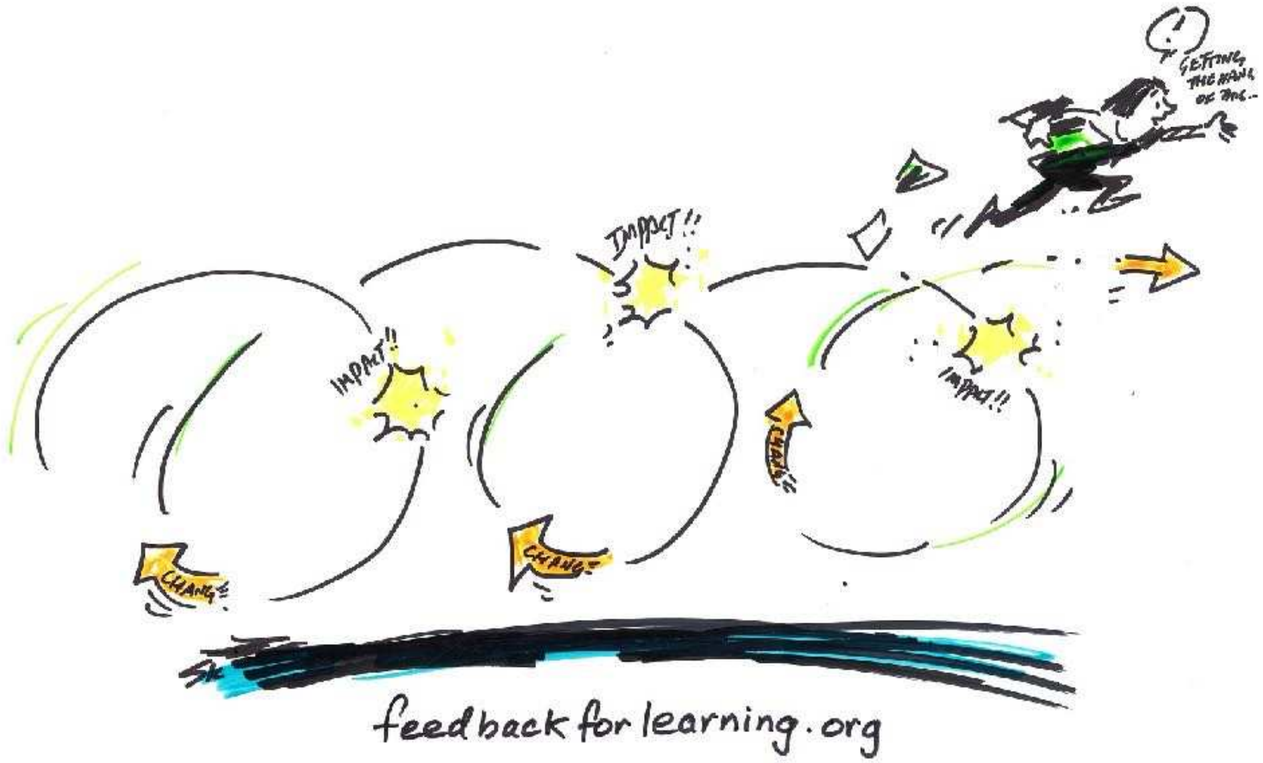
# AI Ethicality: Summary of the ‘WISH LEVEL’



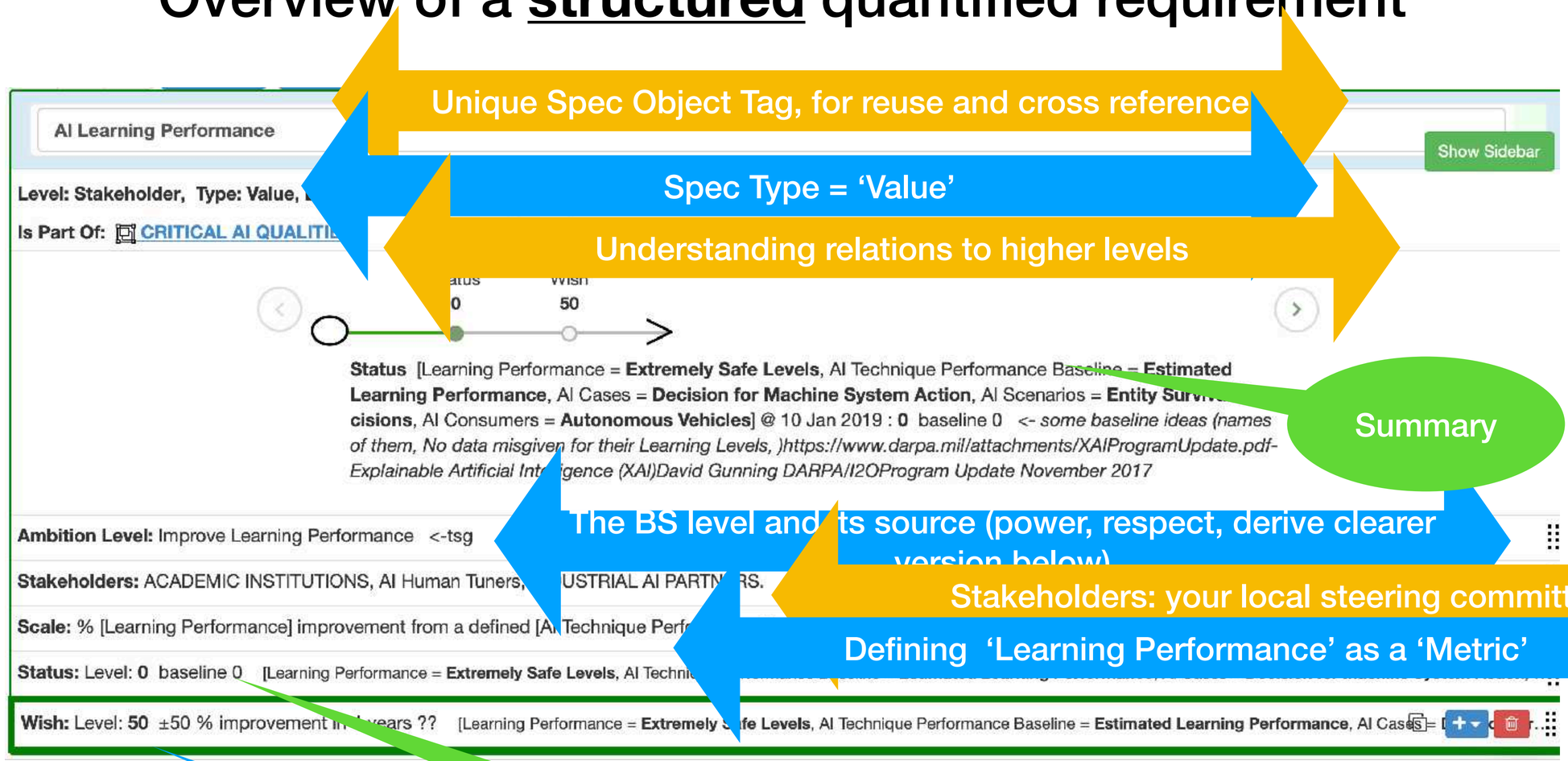

**Wish** [Respect Violation Type = **Legal Violation,Standards Violation,Negative PR Violation**, AI Consumers = **Government,Industry,Commerce,Autonomous Vehicles**, AI Scenarios = **Health Decisions,Entity Survival Decisions**, AI Suppliers = **All**, Market Area = **Core Initial Market**] @ 09 Jan 2027 : **90** ±9 )) *<- tsg example only not real*

Here is an example of selecting a *prioritized*, critical, part of the action,
maybe in practice 1/50 of total budget, and 1/50 total time to deadline (a ‘sprint’ if you like)
Value delivered early, design experiments possible, feedback and correction possible quickly
FREE GIFT REVIEW COPY FOR YOU ALONE. NO COUPON CODE REQUIRED.

# AI Learning Performance



# AI Learning Performance:
# Overview of a **structured** quantified requirement




Unique Spec Object Tag, for reuse and cross reference

Spec Type = ‘Value’

Understanding relations to higher levels

Summary

The BS level and its source (power, respect, derive clearer version below)

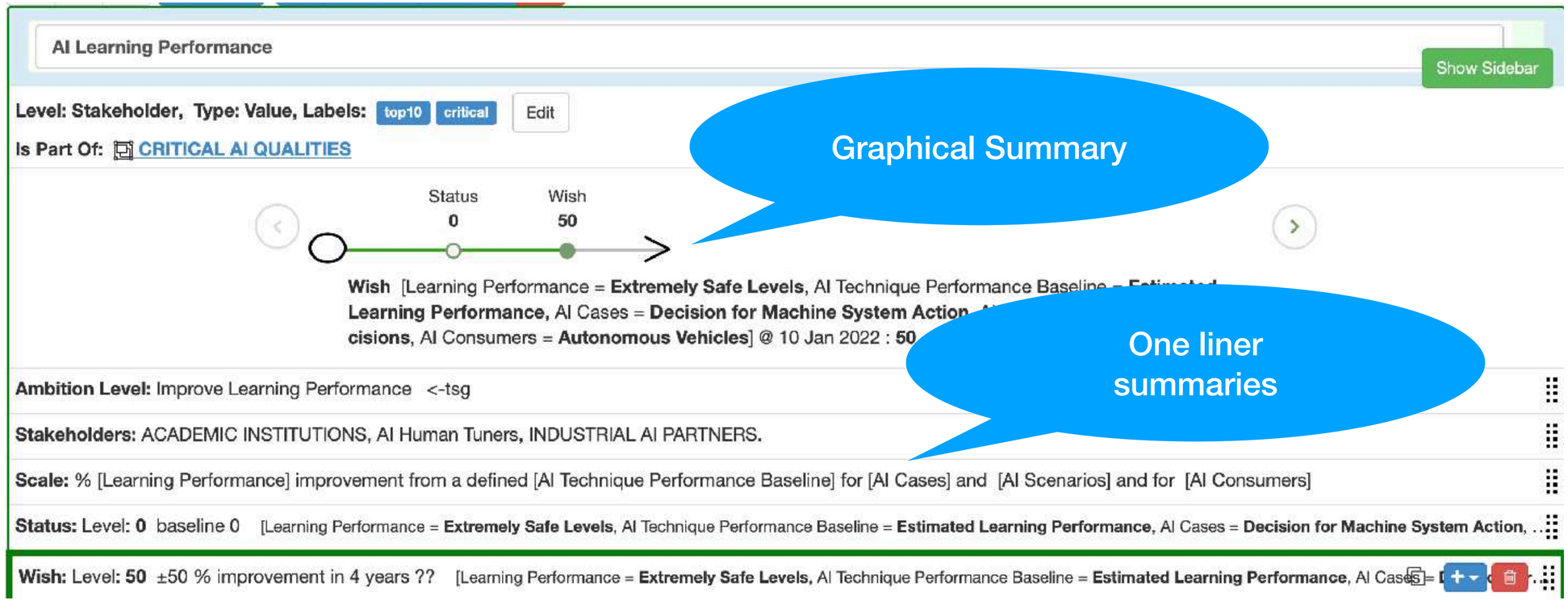
Stakeholders: your local steering committee

Defining ‘Learning Performance’ as a ‘Metric’

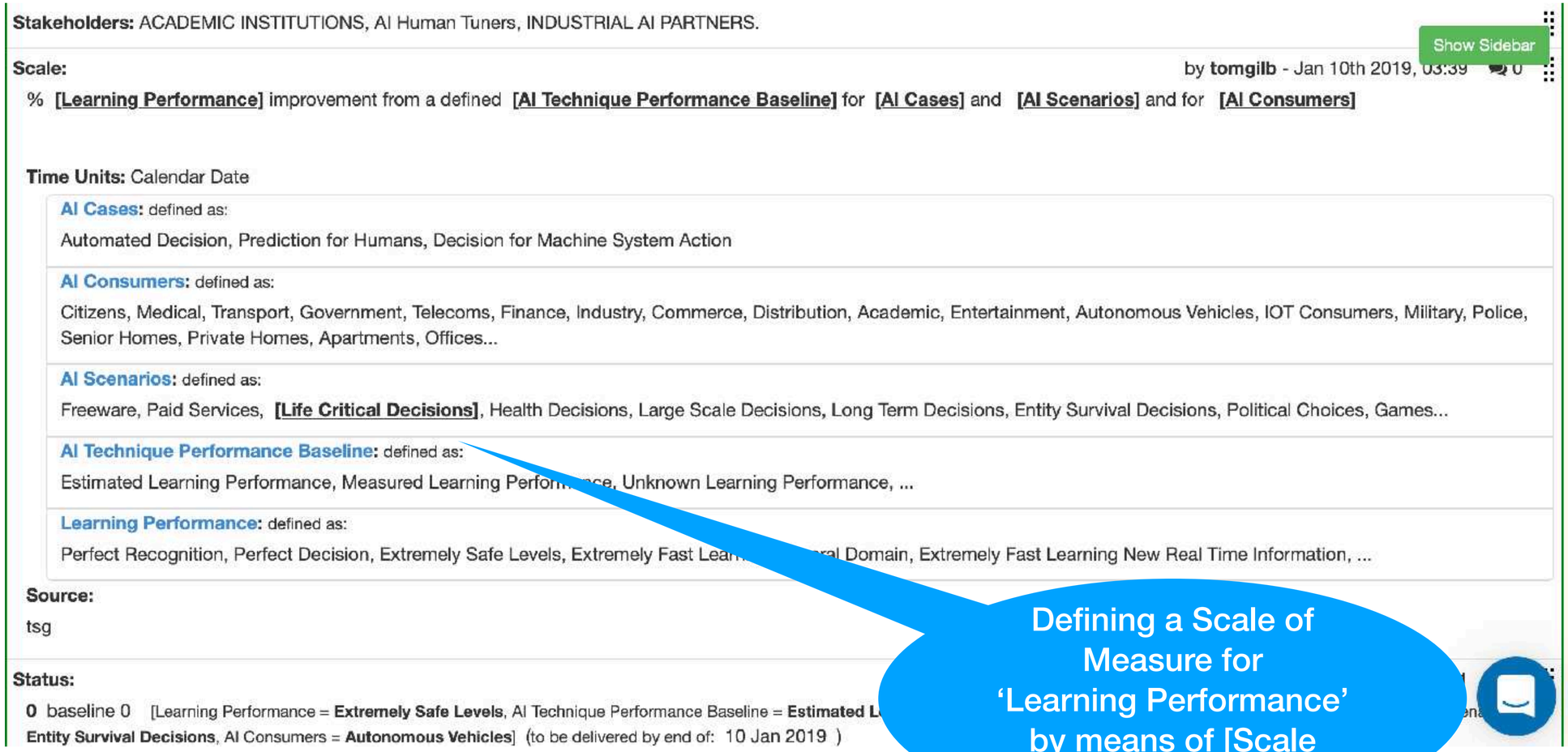
Baseline, systems analysis

Target, acknowledging stakeholder value

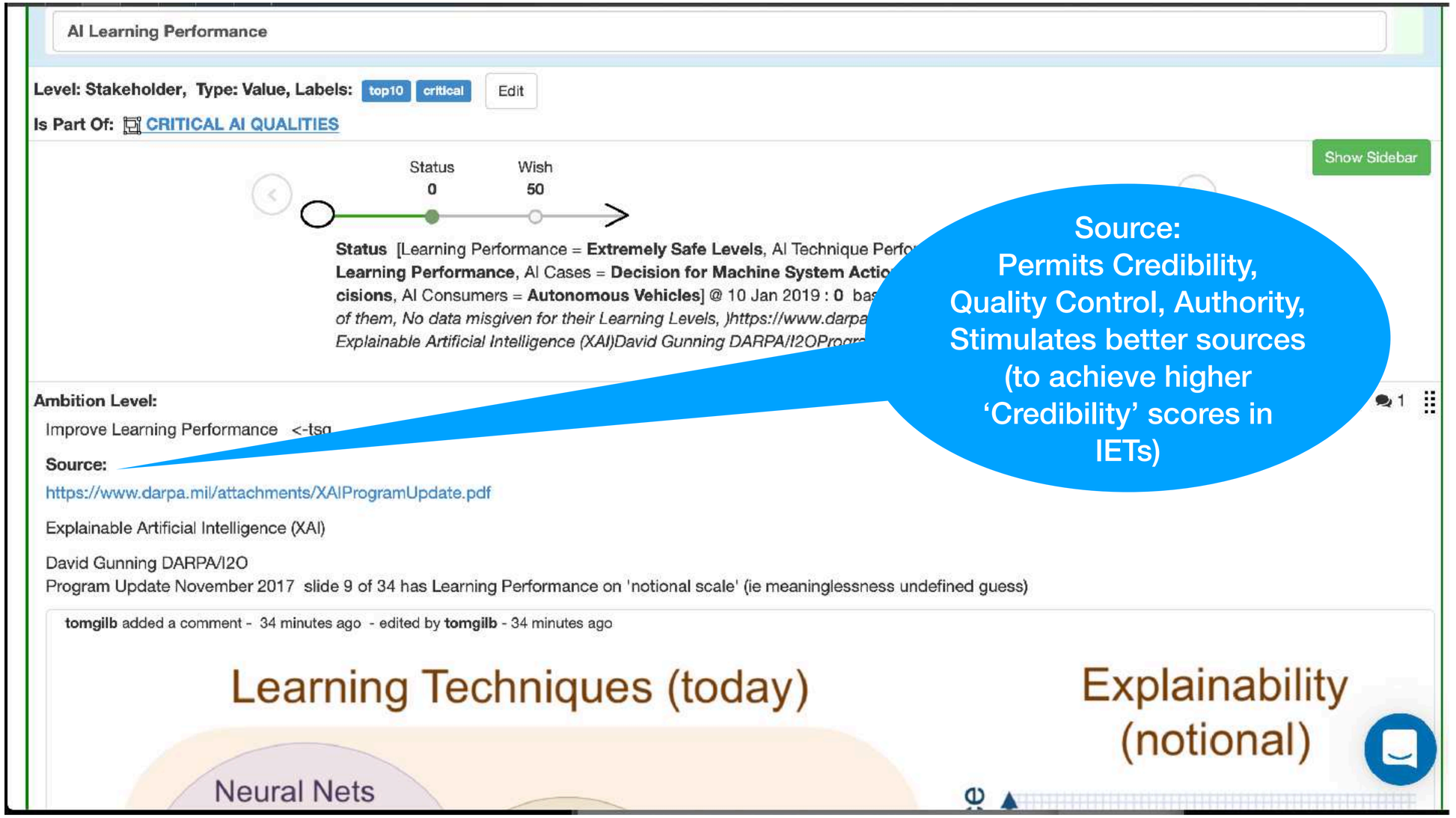
# AI Learning Performance: Wish Level



# AI Learning Performance: Status Level



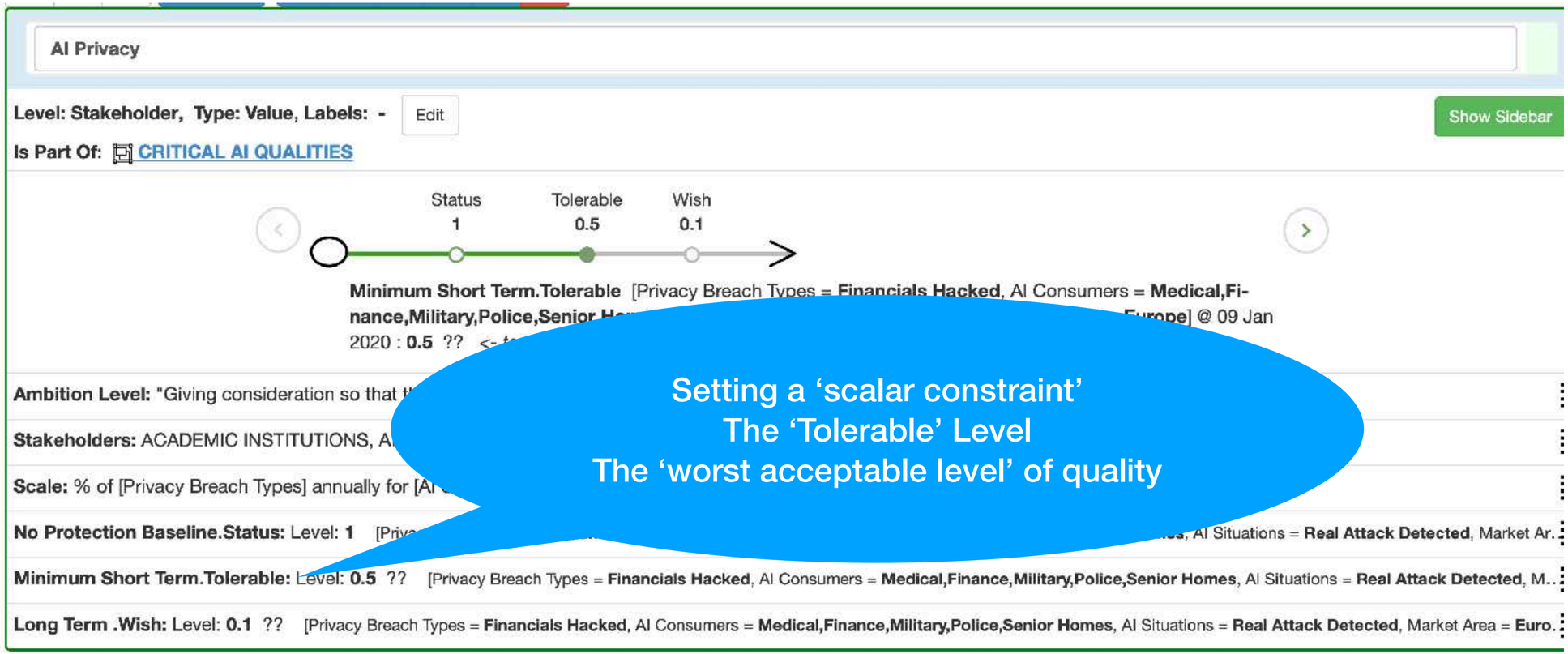
# AI Learning Performance:
# Darpa Source Level Spec: example of source.



# AI Privacy

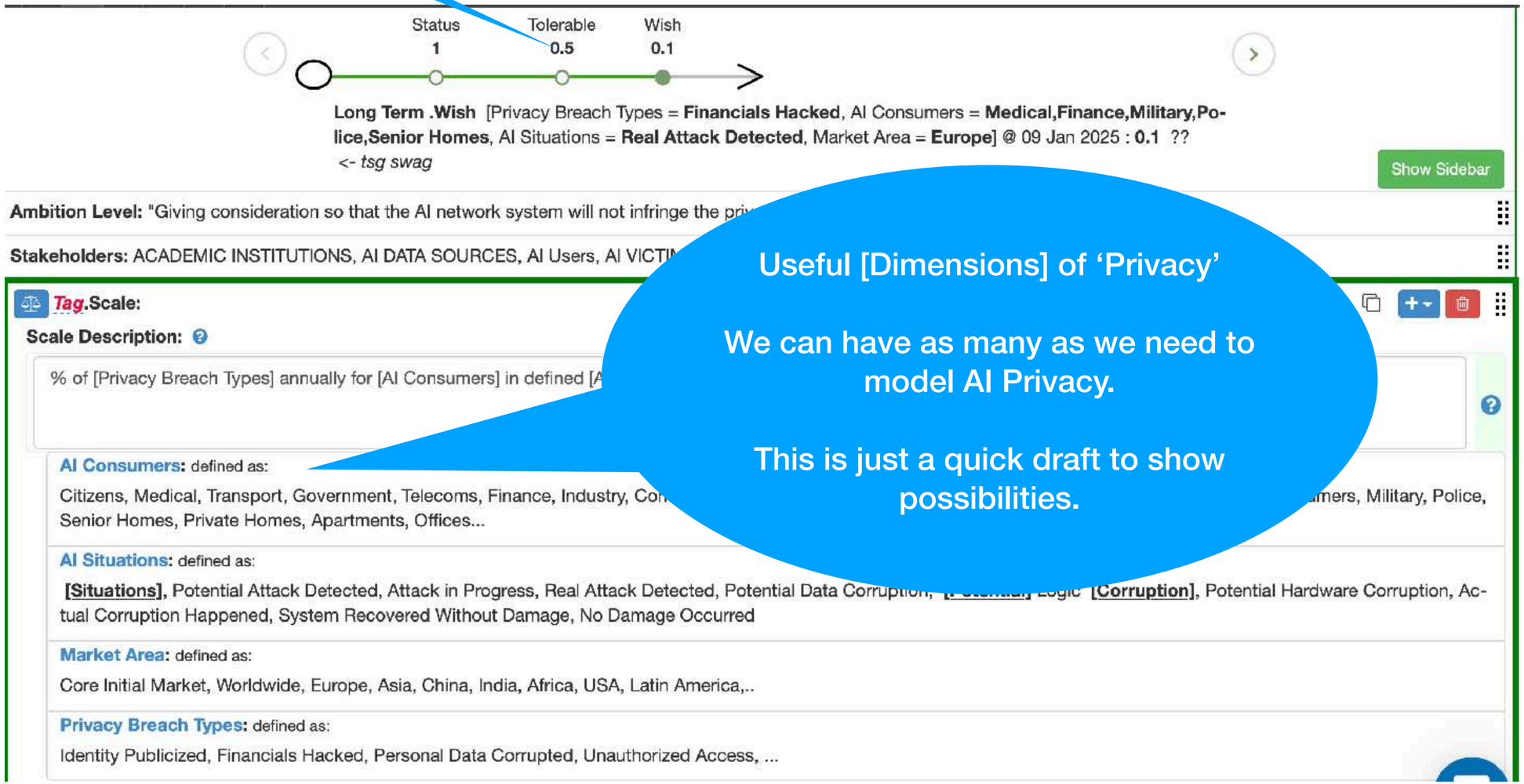

COLLEGE
SOCIAL SECURITY NUMBER
CITY OF BIRTH
ADDRESS
DISABILITY
DATE OF BIRTH
ARREST
SALARY
DRIVER'S LICENSE
DIPLOMA
SCHOOL
AGE
GED
JOB
BANK ACCOUNT
PASSPORT NUMBER
TELEPHONE NUMBER
GENDER

# AI Privacy


AI Privacy
Level: Stakeholder, Type: Value, Labels: -
Edit
Show Sidebar
Is Part Of: CRITICAL AI QUALITIES
Status
1
Tolerable
0.5
Wish
0.1
Minimum Short Term.Tolerable [Privacy Breach Types = Financials Hacked, AI Consumers = Medical,Finance,Military,Police,Senior Homes, AI Situations = Real Attack Detected, Market Area = Europe] @ 09 Jan 2020 : 0.5 ??
Ambition Level: "Giving consideration so that
Stakeholders: ACADEMIC INSTITUTIONS, A
Scale: % of [Privacy Breach Types] annually for [A
No Protection Baseline.Status: Level: 1 [Priv
AI Situations = Real Attack Detected, Market Ar.
Minimum Short Term.Tolerable: Level: 0.5 ?? [Privacy Breach Types = Financials Hacked, AI Consumers = Medical,Finance,Military,Police,Senior Homes, AI Situations = Real Attack Detected, M..
Long Term .Wish: Level: 0.1 ?? [Privacy Breach Types = Financials Hacked, AI Consumers = Medical,Finance,Military,Police,Senior Homes, AI Situations = Real Attack Detected, Market Area = Euro.
Setting a ‘scalar constraint’
The ‘Tolerable’ Level
The ‘worst acceptable level’ of quality

Tolerable

# AI Privacy




Useful [Dimensions] of ‘Privacy’

We can have as many as we need to model AI Privacy.

This is just a quick draft to show possibilities.

# AI Safety

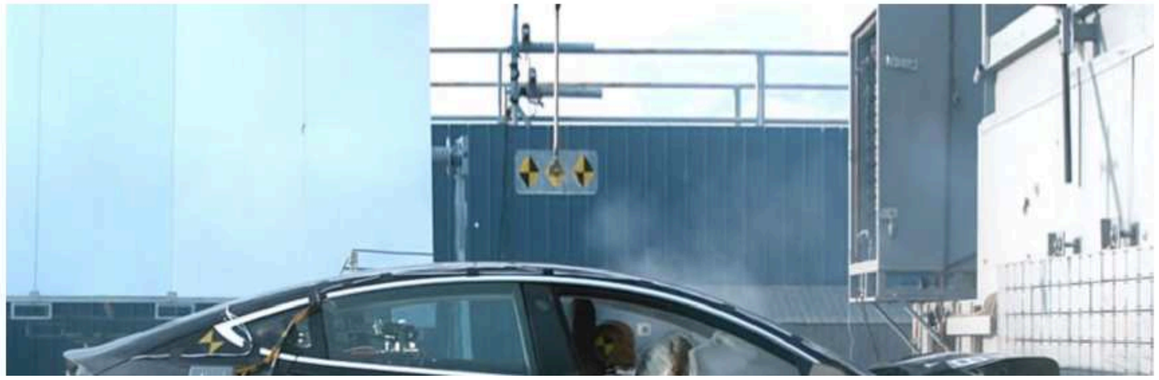
## According To Tesla, The Model 3 Is The Safest Car On The Planet

OCT 9, 2018 / BY JAY TRAUGOTT IN TESTING

### And the number two and three spots go to...

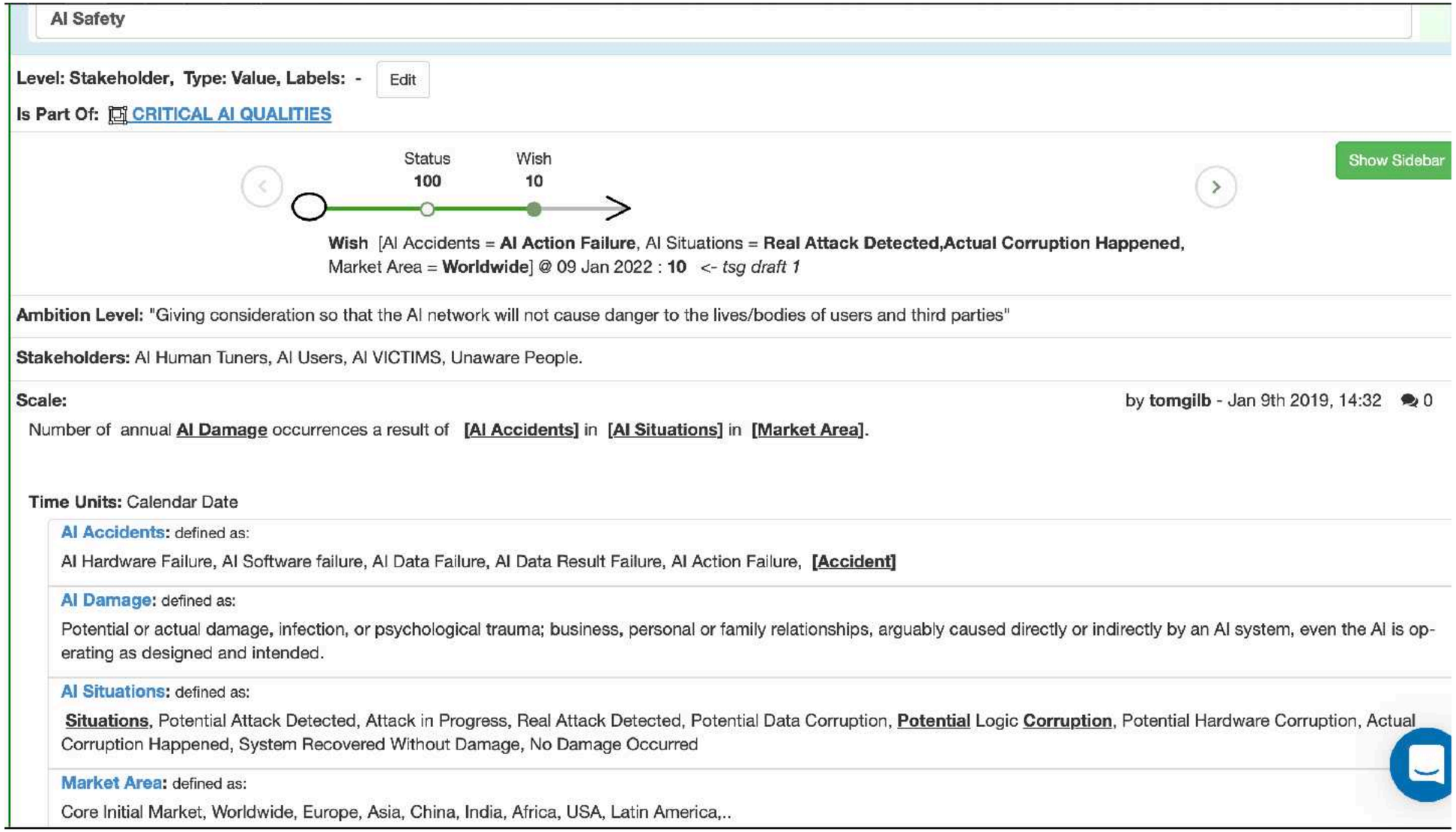
Last month it was announced the new Tesla Model 3 earned a perfect five-star crash test rating in every category from the National Highway Transportation Safety Administration. Most impressive. And now Tesla says its new mid-size sedan has the lowest probability of causing injury for any vehicle ever tested by the US government agency.

"Not only has Model 3 achieved a perfect 5-star safety rating in every category and sub-category, but NHTSA's tests also show that it has the lowest probability of injury of all cars the safety agency has ever tested," Tesla wrote in a blog post. This same post also explains why the Model 3 is so safe.



# AI Safety

## Is your family safety worth quantifying?
## or is ‘Make My Family Safe Again’ Good enough ?

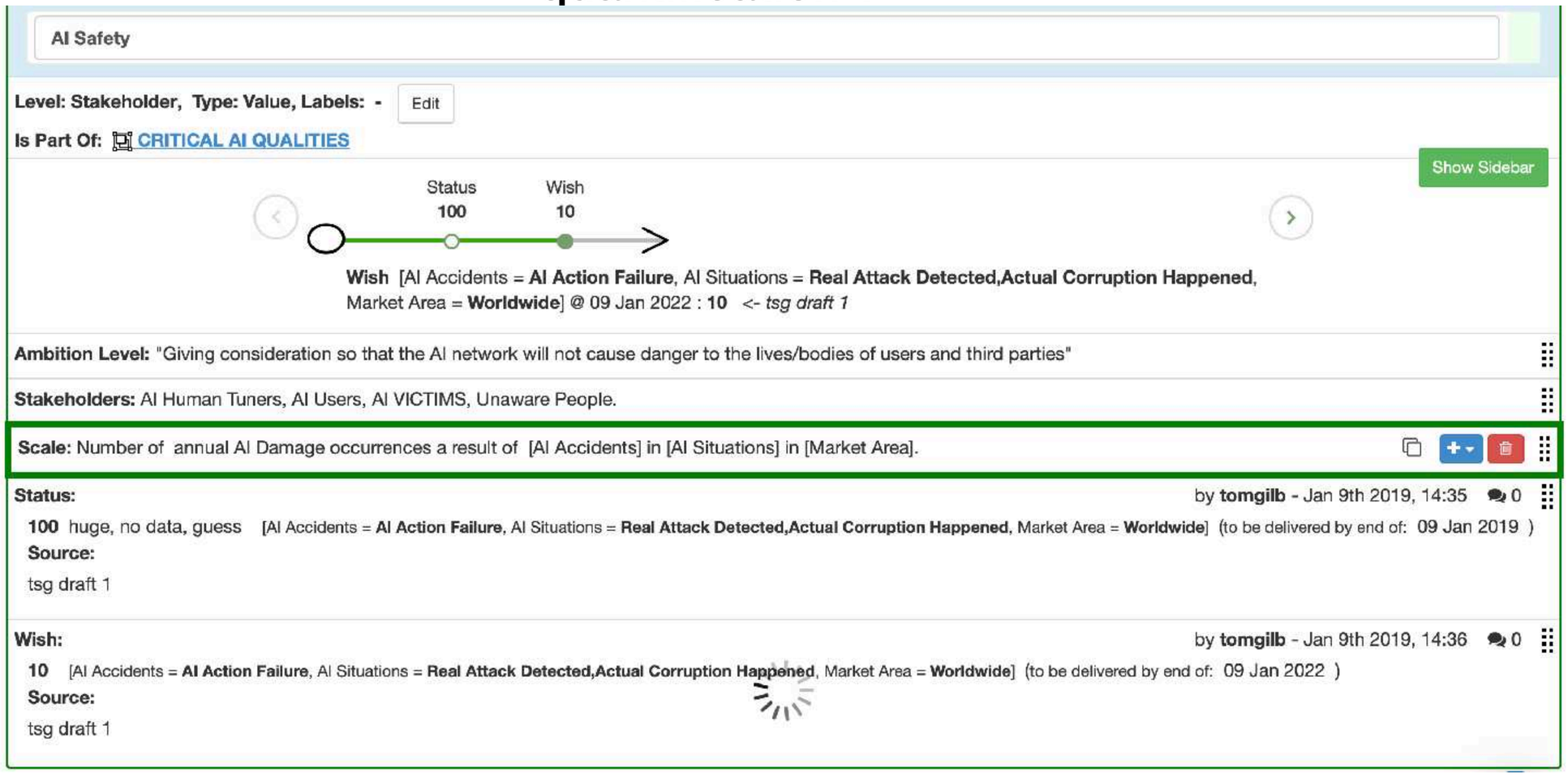


# AI Safety

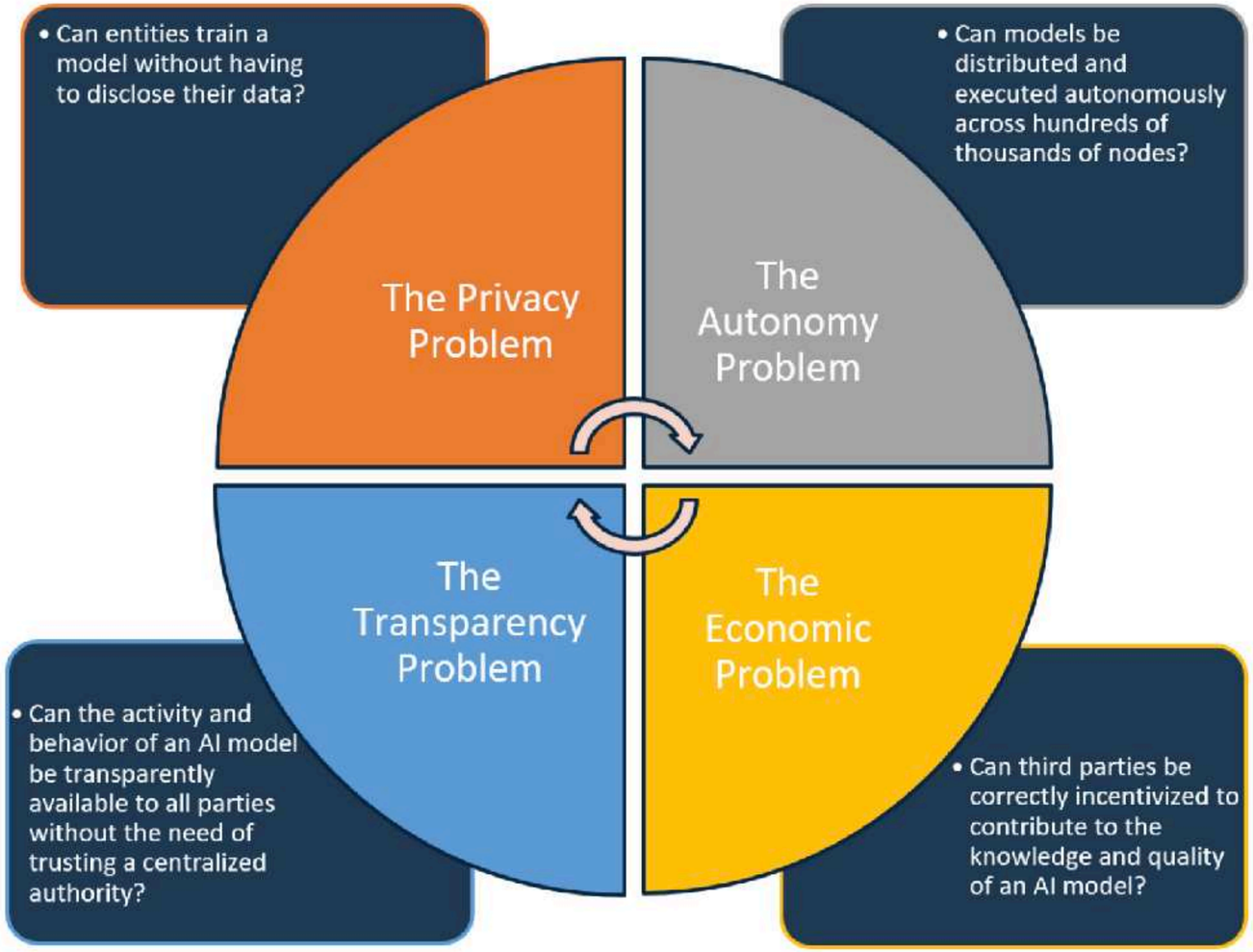
So here is an example of quantification of Safety.

Why don't the AI Standards groups INSIST on quantification of Critical AI Attributes?

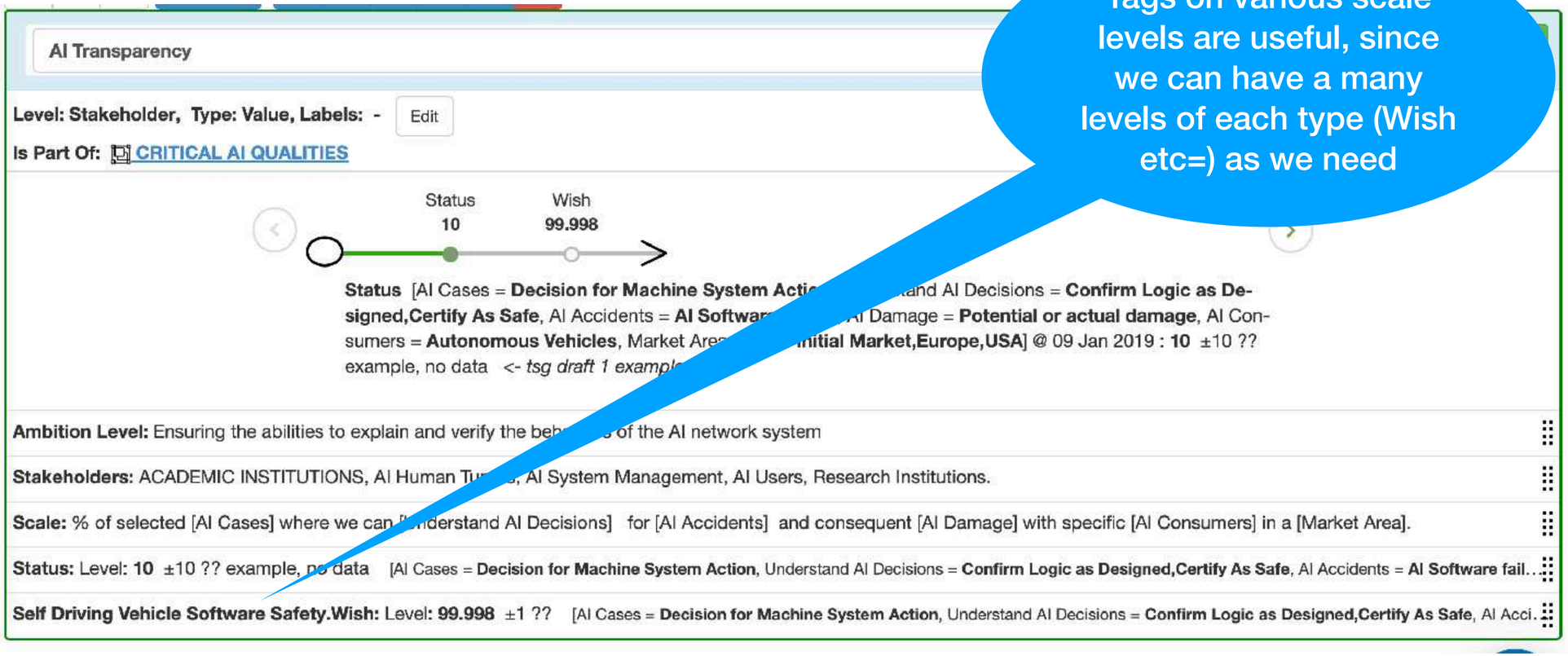
Maybe… they do not intellectually understand 'quantification'?



# AI Transparency



# AI Transparency




Tags on various scale levels are useful, since we can have a many levels of each type (Wish etc=) as we need

# AI Transparency

**Stakeholders:** ACADEMIC INSTITUTIONS, AI Human Tuners, AI System Management, AI Users, Research Institutions.

**Scale:** by **tomgilb** - Jan 9th 2019, 15:11 0

% of selected **[AI Cases]** where we can **[Understand AI Decisions]** for **[AI Accidents]** and consequent **[AI Damage]** with specific **[AI Consumers]** in a **[Market Area]**.

Do you remember the fuzzy Ambition Level previous slide ?

Useless: shame for doing it in public!

**Time Units:** Calendar Date

**AI Accidents:** defined as:

AI Hardware Failure, AI Software failure, AI Data Failure, AI Data Result Failure, AI Actic

**AI Cases:** defined as:

Automated Decision, Prediction for Humans, Decision for Machine System Action

**AI Consumers:** defined as:

Citizens, Medical, Transport, Government, Telecoms, Finance, Industry, Commerce, Distribution, Academic, Entertainment, Autonomous Vehicles, IOT Consumers, Military, Police, Senior Homes, Private Homes, Apartments, Offices...

**AI Damage:** defined as:

Potential or actual damage, infection, or psychological trauma; business, personal or family relationships, arguably caused directly or indirectly by an AI system, even the AI is operating as designed and intended.

**Market Area:** defined as:

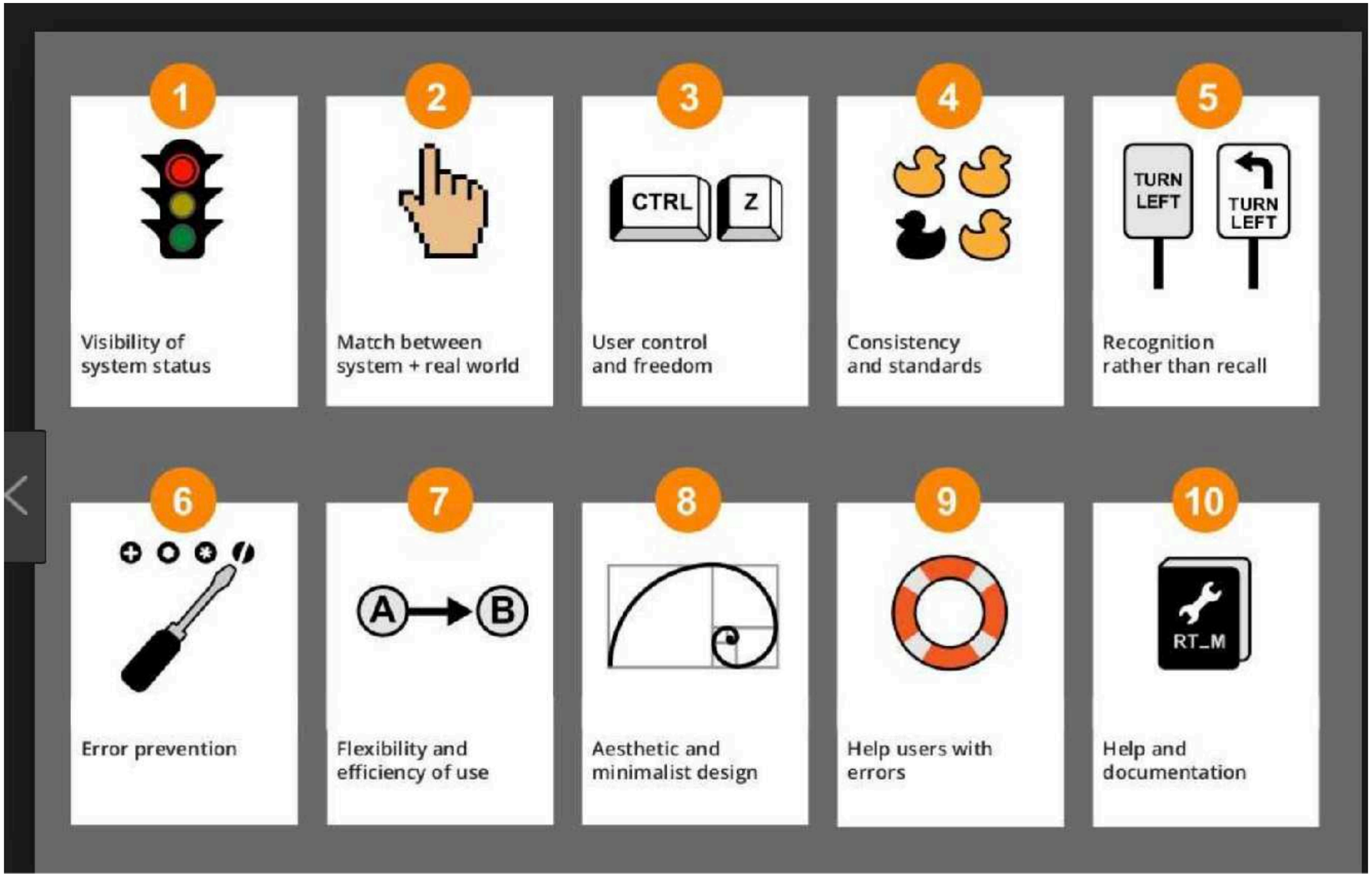
Core Initial Market, Worldwide, Europe, Asia, China, India, Africa, USA, Latin America,..

**Understand AI Decisions:** defined as:

Explain Logic, Know Data Used, Confirm Logic as Designed, Certify As Safe, Regression Test After Changes, Display Detailed Logic Path for Analysis, ...

**Status:** Level: **10** ±10 ?? example, no data [AI Cases = **Decision for Machine System Action**, Understand AI Decisions = **Confirm Lo**

# AI Usability



# AI Usability

**Ambition Level:** "Giving consideration so that the AI network system can assist users and appropriately provide users with opportunities to make choices."

**Stakeholders:** by **tomgilb** - Jan 9th 2019, 01:45 0

| Stakeholder | Roles | Notes |
|---|---|---|
| From: AI Users | | |
| From: Aware People | | |

Show Sidebar

**Scale:** by **tomgilb** - Jan 9th 2019, 15:38 0

Speed in Minutes for an **[AI User Type]** to **Successfully Use AI** for **[AI Scenarios]** in a **[Market Area]**.

**Time Units:** Calendar Date

**AI Scenarios:** defined as:

Freeware, Paid Services, **[Life Critical Decisions]**, Health Decisions, Large Scale Decisions, Long Term Decisions, Entity Survival Decisions, Political Choices, Games...

**AI User Type:** defined as:

Unaware of AI, Conscious of AI,

Beginner, Self Taught, Safe User, Expert Level

**Market Area:** defined as:

Core Initial Market, Worldwide, Europe, Asia, China, India, Africa, USA, Latin America,..

**Successfully Use AI:** defined as:

Use AI as intended, without delays or faults.

**Source:**

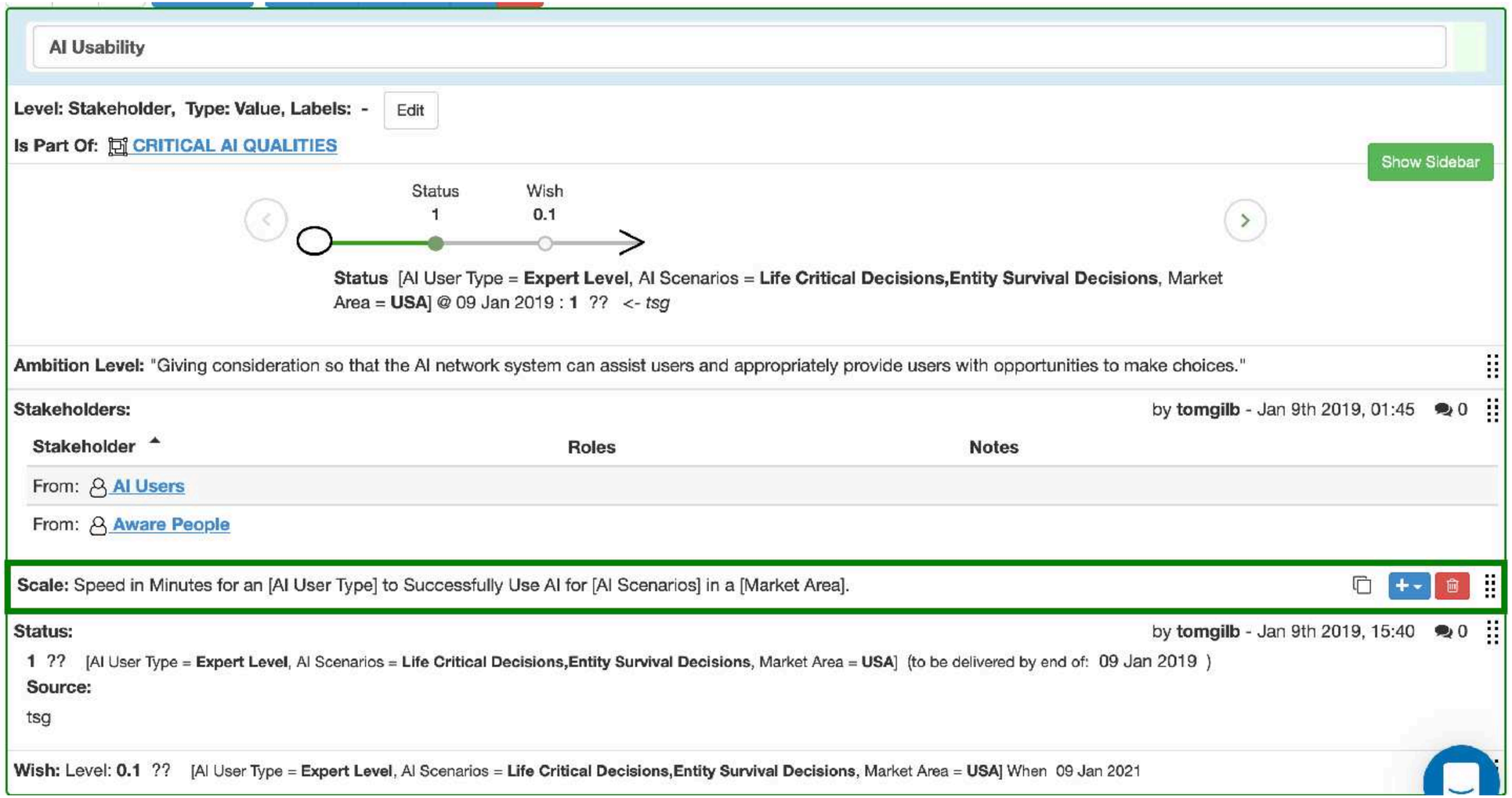
tsg draft 1

**Status:** by **tomgilb** - Jan 9th 2019, 15:40

**1** ?? [AI User Type = **Expert Level**, AI Scenarios = **Life Critical Decisions,Entity Survival Decisions**, Market Area = **USA**] (to be delivered by end of: 09 Jan 2019 )

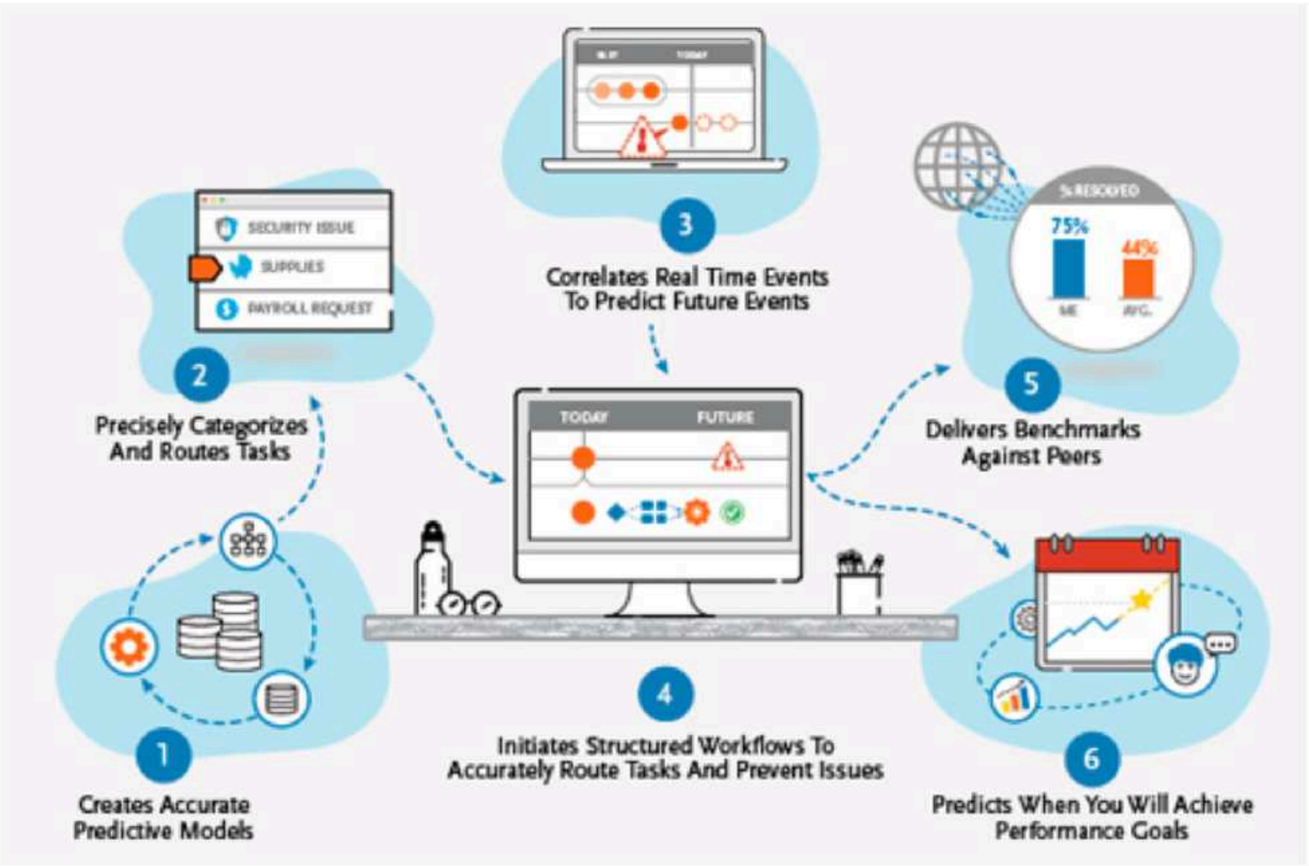
**Source:**

# AI Usability



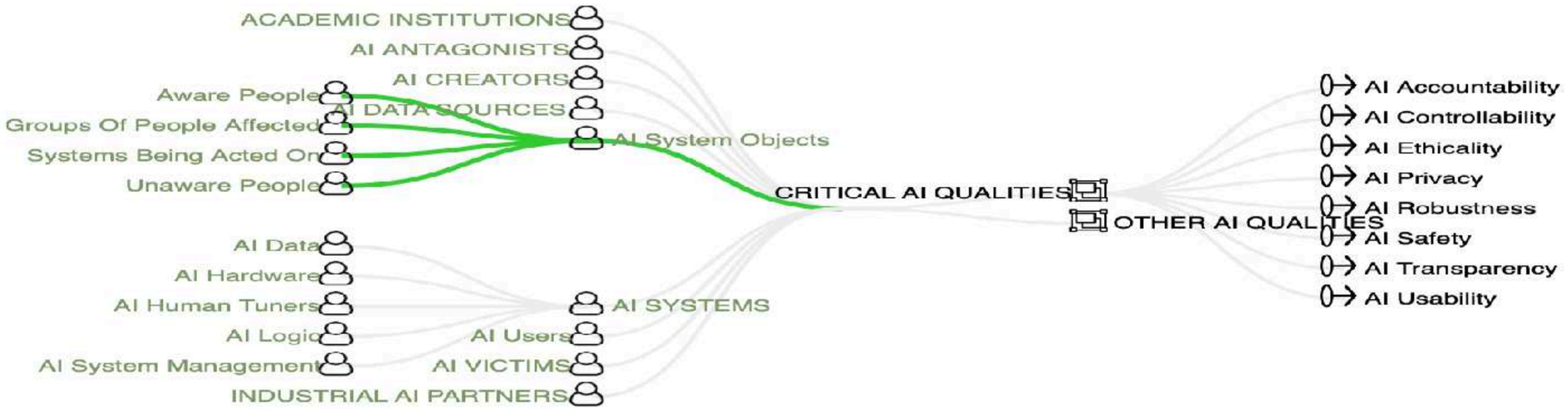
# Other Viewpoints

2020.




Source: https://www.businesswire.com/news/home/20170509005837/en/ServiceNow- Launches-Intelligent-Automation-Engine%E2%84%A2

# Diagram: AI Stakeholders & AI ‘Qualities’ (Principles)

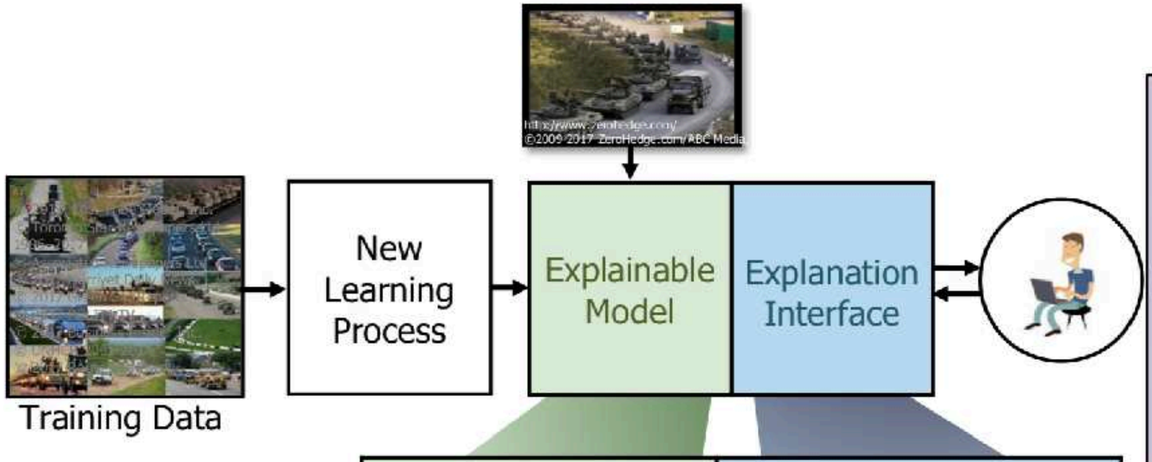


# Diagram:
# Stakeholders, Qualities,
# Some ‘XAI Techniques’

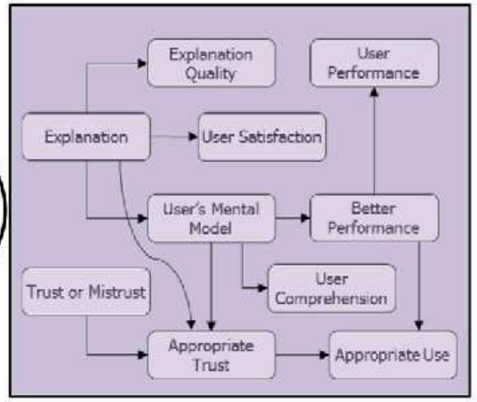
https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

# XAI Concept and Technical Approaches




| | | |
|---|---|---|
| **UC Berkeley** | Deep Learning | Reflexive and Rational |
| **Charles River Analytics** | Causal Modeling | Narrative Generation |
| **UCLA** | Pattern Theory+ | 3-Level Explanation |
| **Oregon State** | Adaptive Programs | Acceptance Testing |
| **PARC** | Cognitive Modeling | Interactive Training |
| **CMU** | Explainable RL (XRL) | XRL Interaction |
| **SRI International** | Deep Learning | Show and Tell Explanations |
| **Raytheon BBN** | Deep Learning | Argumentation and Pedagogy |
| **UT Dallas** | Probabilistic Logic | Decision Diagrams |
| **Texas A&M** | Mimic Learning | Interactive Visualization |
| **Rutgers** | Model Induction | Bayesian Teaching |

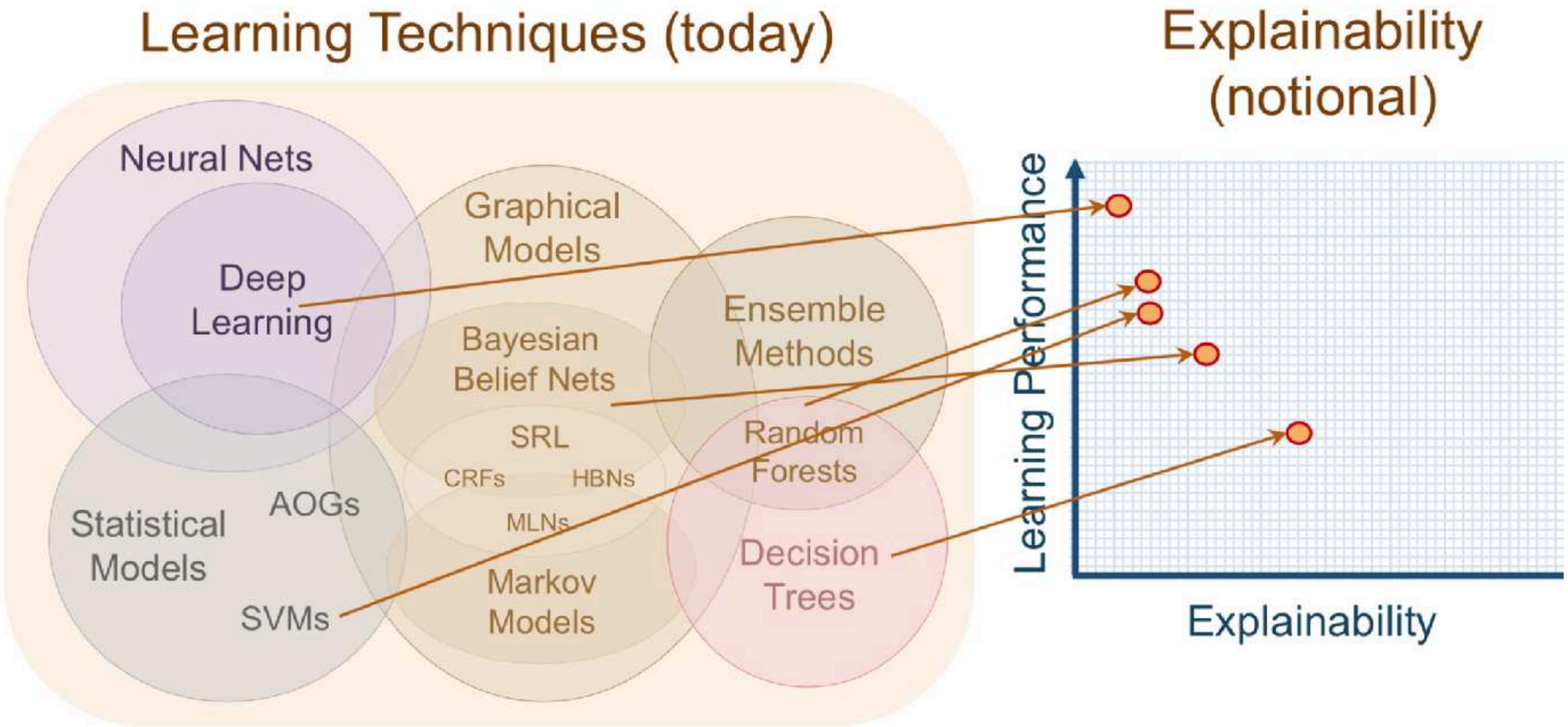



**IHMC**
Psychological Model of Explanation

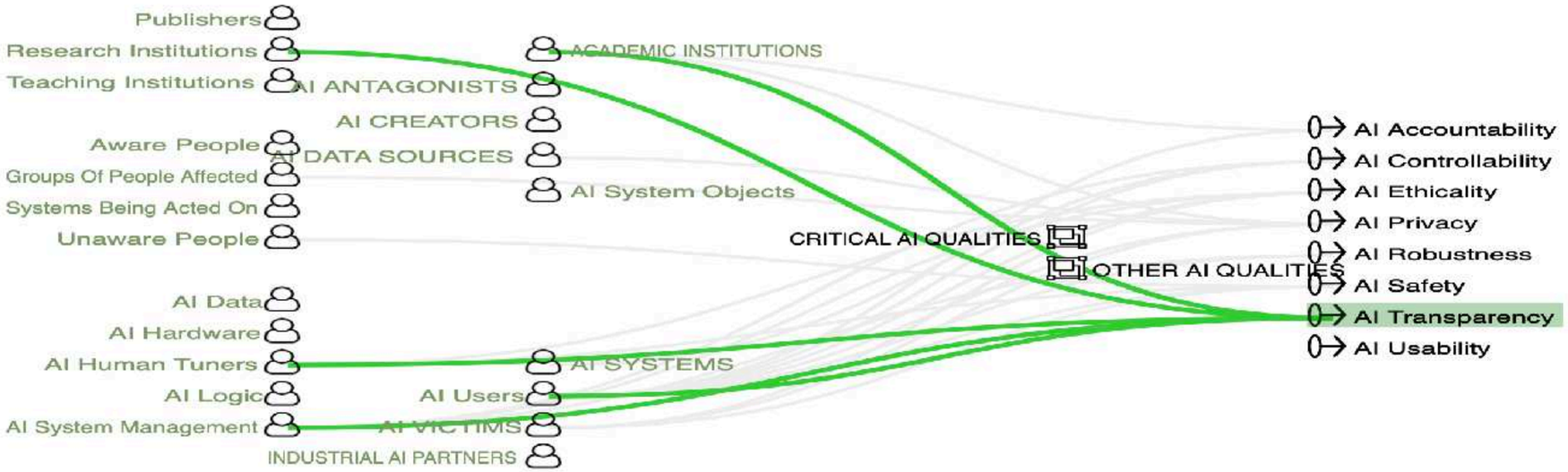
# XAI Developers (TA1)

| CP | Performer | Explainable Model | Explanation Interface |
|---|---|---|---|
| Both | UC Berkeley | Deep Learning | Reflexive and Rational |
| | Charles River | Causal Modeling | Narrative Generation |
| | UCLA | Pattern Theory+ | 3-level Explanation |
| Autonomy | Oregon State | Adaptive Programs | Acceptance Testing |
| | PARC | Cognitive Modeling | Interactive Training |
| | CMU | Explainable RL (XRL) | XRL Interaction |
| Analytics | SRI International | Deep Learning | Show and Tell Explanation |
| | Raytheon BBN | Deep Learning | Argumentation and Pedagogy |
| | UT Dallas | Probabilistic Logic | Decision Diagrams |
| | Texas A&M | Mimic Learning | Interactive Visualization |
| | Rutgers | Model Induction | Bayesian Teaching |

# Explainability of Learning Techniques

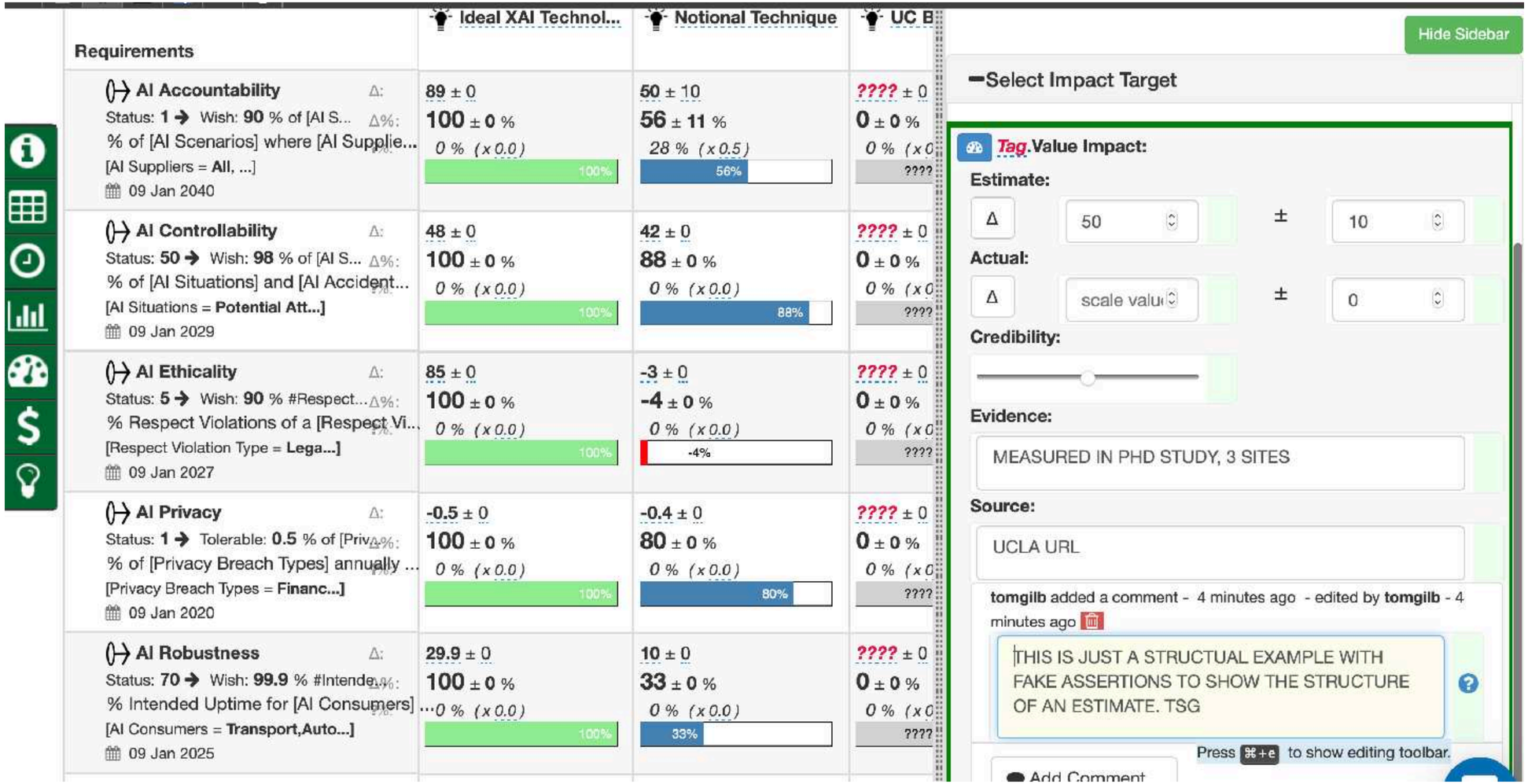


# Relationship Diagram
# AI Transparency and its stakeholders



# Can we evaluate AI Technology Qualities against 9 AI Qualities and some Costs?

# AI Technique Evaluation using DATA with objective evidence, sources, AND uncertainty

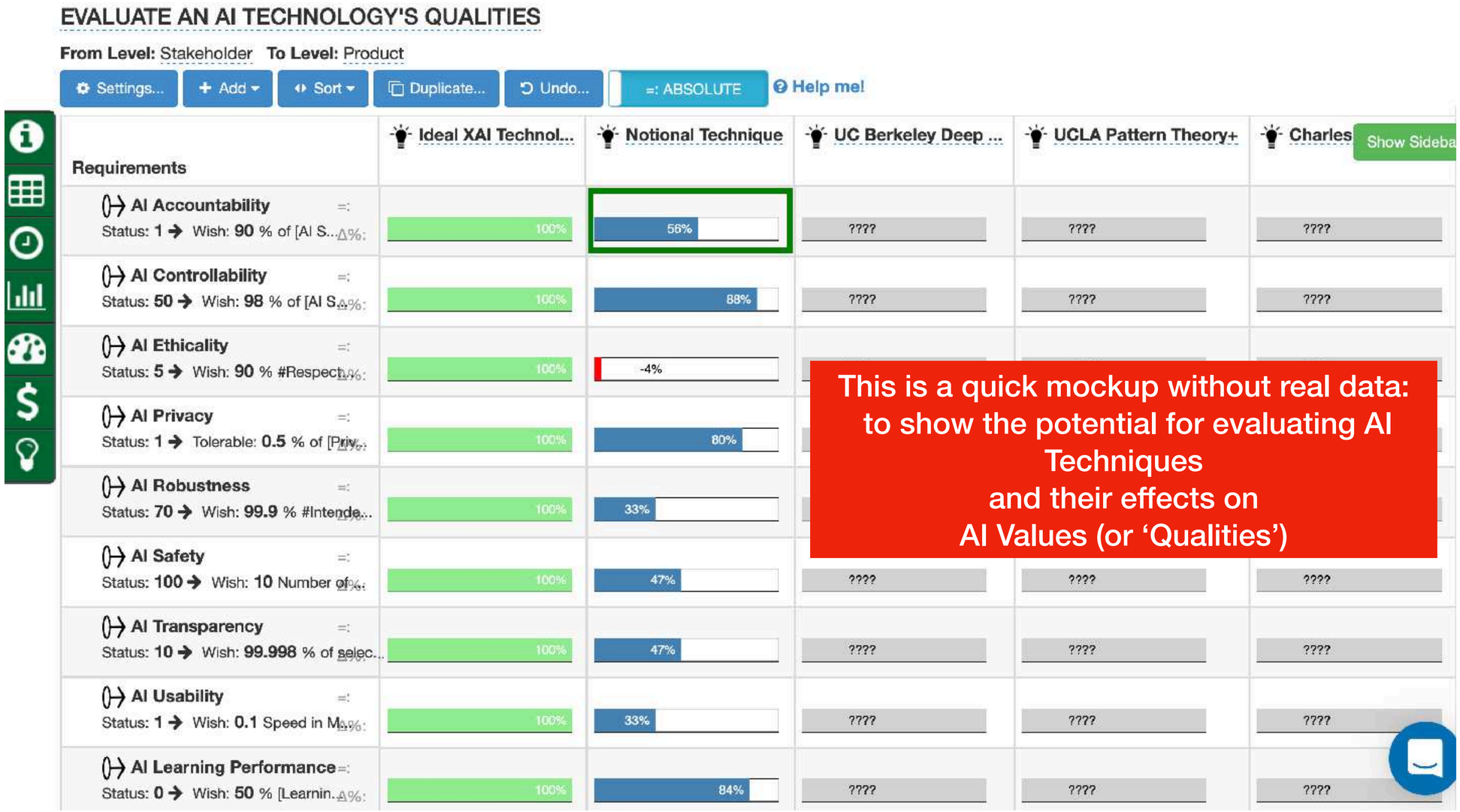



**small diagram problem: the selected window is not shown. 230619 tsg**

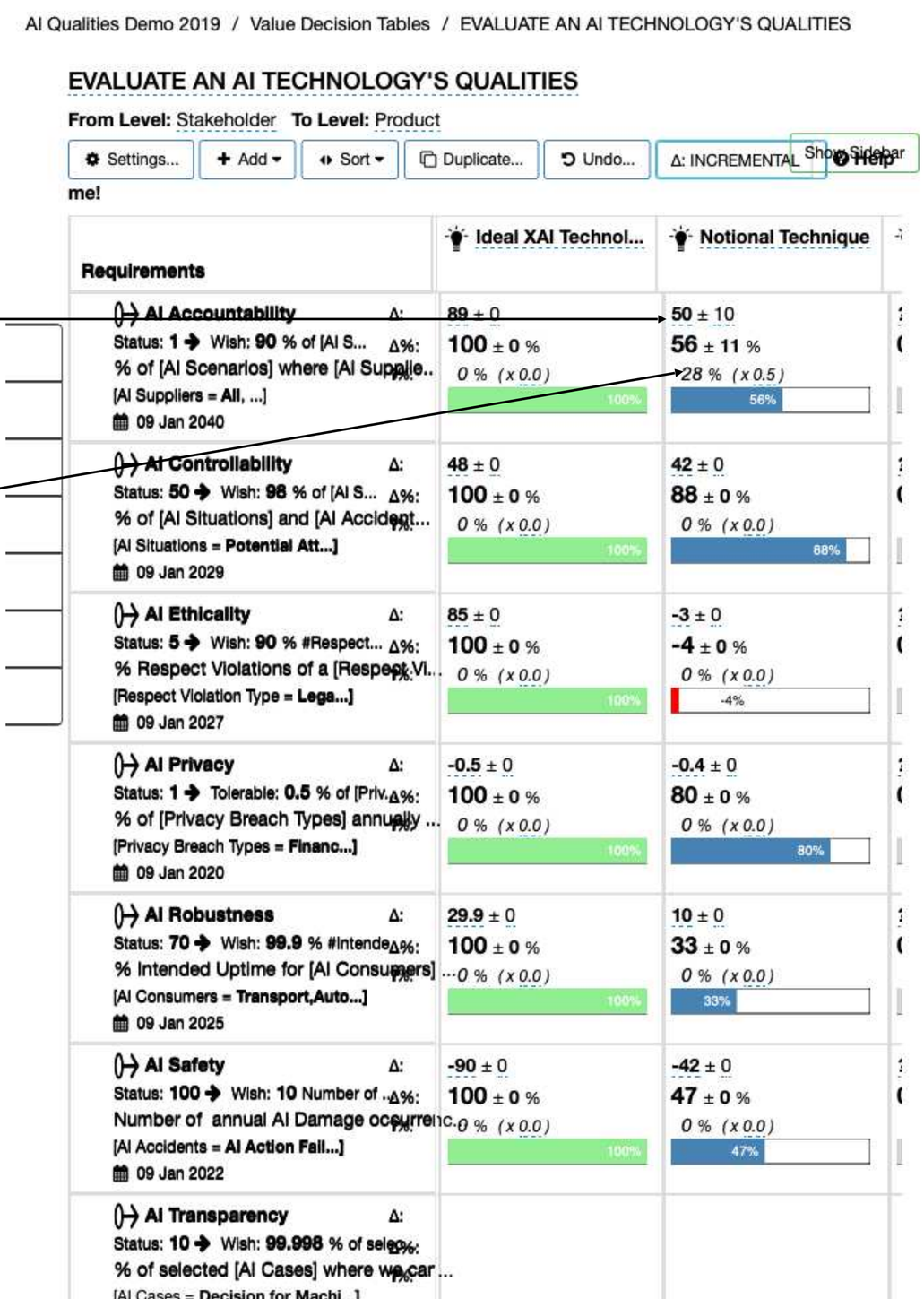
# Defined Concepts: Not least for [Scale Parameters]

| | | | | |
|---|---|---|---|---|
| AI Accidents | AI Hardware Failure, AI Software failure, AI Data Failure, AI Data Result Failure, AI Action Failure, [Accident] | @tomgilb | Jan 9th 2019, 10:00 | tomgilb |
| AI Consumers | Citizens, Medical, Transport, Government, Telecoms, Finance, Industry, Commerce, Distribution, Academic, Entertainment, Autonomous Vehicles, IOT Consumers, Military, Police, Senior Homes, Private Homes, Apartments, Offices... | @tomgilb | Jan 9th 2019, 10:06 | tomgilb |
| AI Management | AI Supplier Service, User Organization Service, User Service, Service Organization | @tomgilb | Jan 9th 2019, 09:59 | tomgilb |
| AI Scenarios | Freeware, Paid Services, [Life Critical Decisions], Health Decisions, Large Scale Decisions, Long Term Decisions, Entity Survival Decisions, Political Choices, ... | @tomgilb | Jan 9th 2019, 06:09 | tomgilb |
| AI Situations | [Situations], Potential Attack Detected, Attack in Progress, Real Attack Detected, Potential Data Corruption, [Potential] Logic [Corruption], Potential Hardware Corruption, Actual Corruption Happened, System Recovered Without Damage, No Damage Occurred | @tomgilb | Jan 9th 2019, 10:00 | tomgilb |
| AI Status | Normal Successful Operation, Dominant Learning Mode, Partial Decision Failures, Frequent Decision Failures, Out of Operation, Recovery Mode, Secondary Backup System Mode, Substantial Human Intervention Necessary | @tomgilb | Jan 9th 2019, 10:00 | tomgilb |
| AI Suppliers | App Developers, Data suppliers, Academic Research, Startups, Consultants, IOT suppliers, Teaching Institutions, ... | @tomgilb | Jan 9th 2019, 06:08 | tomgilb |
| Accountability Strength | Legally Compliant, Contractually Compliant, Standards Compliant, International Guidelines Compliant, ... | @tomgilb | Jan 9th 2019, 06:08 | tomgilb |
| Life Critical Decisions | At least one human life can be lost or effectively destroyed by bad decisions as a result of the AI tool | @tomgilb | Jan 9th 2019, | tomgilb |

# Quantitative Evaluation of AI Technology Qualities using ‘Impact Estimation Tables’



A more detailed example of rating ‘uncertainty’ (±10) and Credibility 56% x 0.5 = 28%

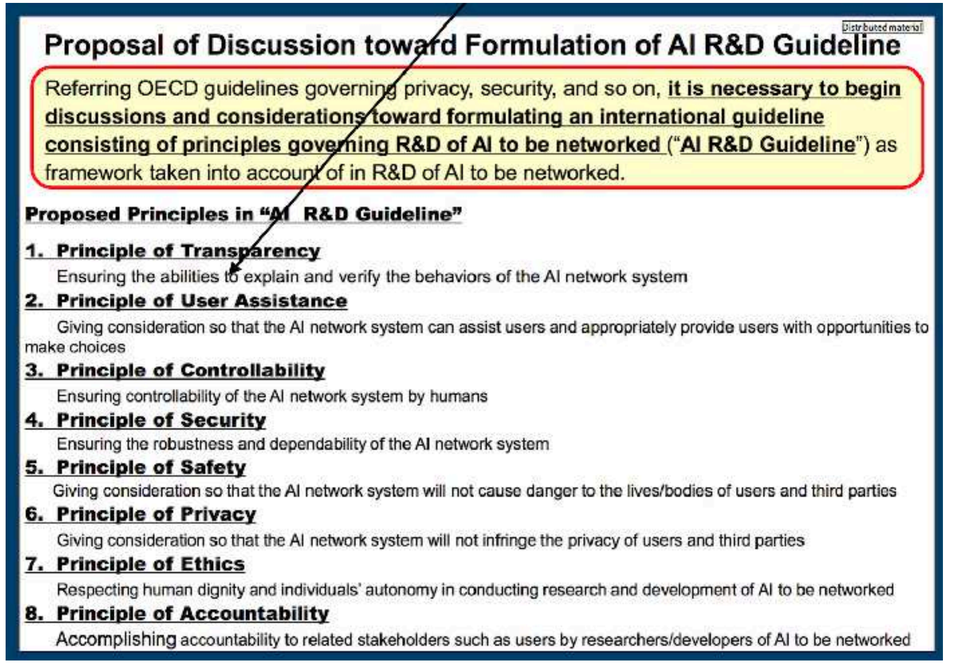



# Mock example

# My (TSG) Observations and Conclusions

- The most fundamental obstacle to AI Standards progress is WE NEED TO **QUANTIFY** AND STRUCTURE-**RICHLY** all **CRITICAL** STAKEHOLDER **VALUES (ALL > 8)**

- Next: we need much-deeper, more-comprehensive identification, and consensus, and detailed knowledge (their values and constraints) about AI/ AI STANDARDS **STAKEHOLDERS**

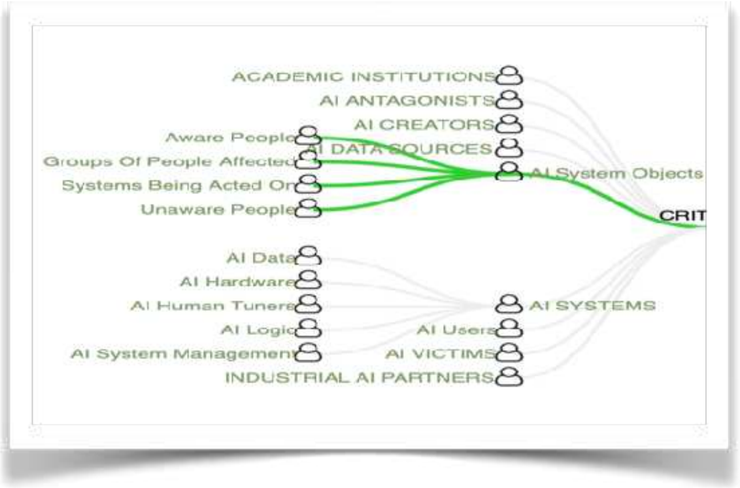

Distributed material

**Proposal of Discussion toward Formulation of AI R&D Guideline**

Referring OECD guidelines governing privacy, security, and so on, **it is necessary to begin discussions and considerations toward formulating an international guideline consisting of principles governing R&D of AI to be networked** ("**AI R&D Guideline**") as framework taken into account of in R&D of AI to be networked.

**Proposed Principles in "AI R&D Guideline"**

1. **Principle of Transparency**
   Ensuring the abilities to explain and verify the behaviors of the AI network system
2. **Principle of User Assistance**
   Giving consideration so that the AI network system can assist users and appropriately provide users with opportunities to make choices
3. **Principle of Controllability**
   Ensuring controllability of the AI network system by humans
4. **Principle of Security**
   Ensuring the robustness and dependability of the AI network system
5. **Principle of Safety**
   Giving consideration so that the AI network system will not cause danger to the lives/bodies of users and third parties
6. **Principle of Privacy**
   Giving consideration so that the AI network system will not infringe the privacy of users and third parties
7. **Principle of Ethics**
   Respecting human dignity and individuals' autonomy in conducting research and development of AI to be networked
8. **Principle of Accountability**
   Accomplishing accountability to related stakeholders such as users by researchers/developers of AI to be networked



# End Slide

"In my day, kids didn't build massive, ransomware-spewing botnets. They got a paper route."

brianmooredraws.com

SLIDES FOLDER

DOI to slides pdf: https://doi.org/10.13140/RG.2.2.10935.12960