A Guide to the Dynamics of AI Principles:

Making Sense of AI Principles & How to Use Them

AI ETHICS LAB

February 2020

last update: June 2020

A Guide to the **Dynamics of AI Principles**

This guide will help you:

- learn more about our toolbox, *Dynamics of AI Principles*,
- understand the structure behind AI principles, and
- systematize & operationalize the AI principles.

Let's get started! 🏁

Who Should Use this Guide?

LEADERS who are responsible for setting the Ethics Strategy of their companies to mitigate and proactively address AI ethics risks.

CREATORS who make ethically loaded decisions in their everyday practice as they build new AI technologies.

ANYONE who wants to understand the function of AI principles.

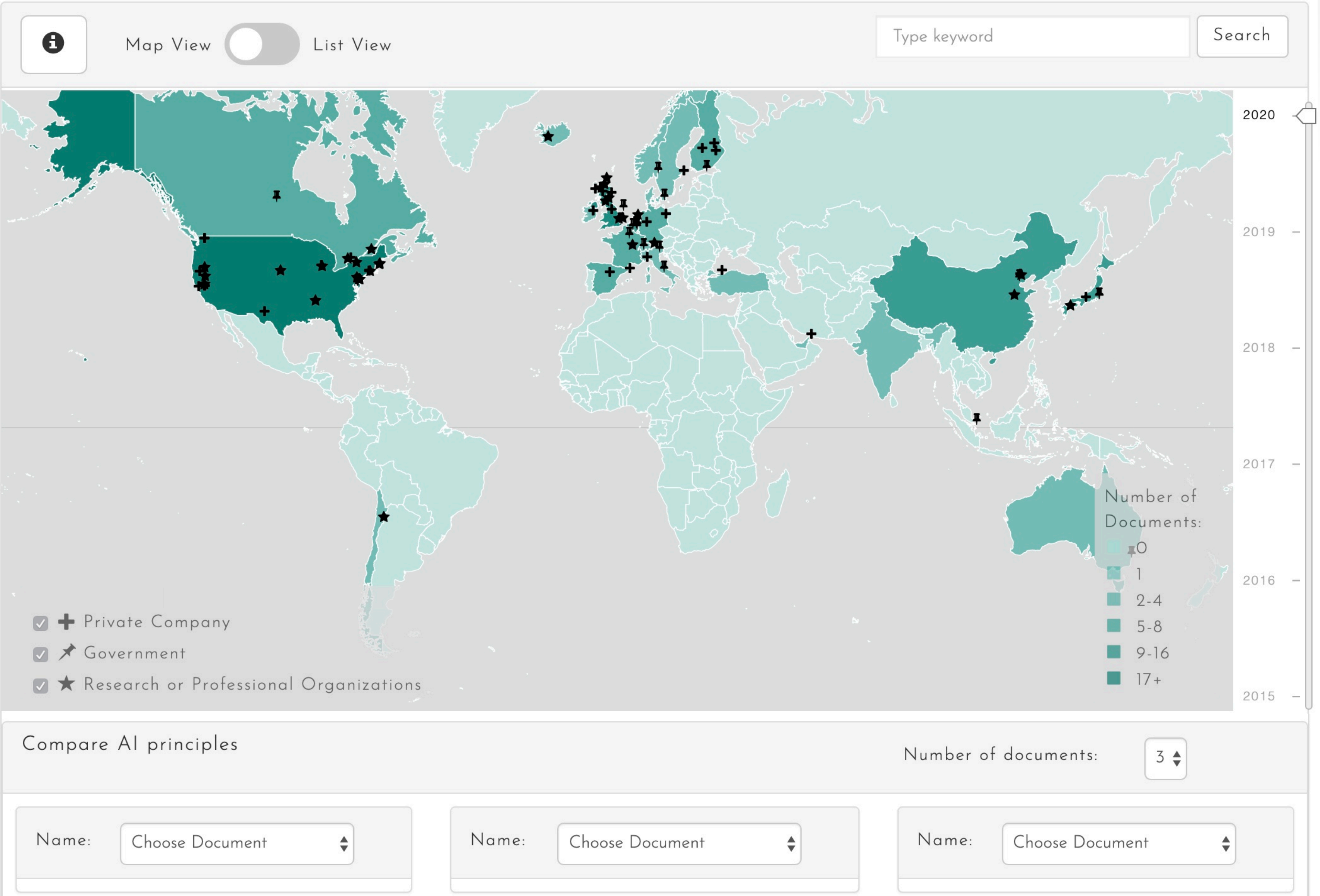
What is the Dynamics of AI Principles?

The **Dynamics of AI Principles** is our toolbox for keeping track of, systematizing, and operationalizing the bewildering and growing number of AI Principles out there.

AI Principles are published by private companies, governmental agencies, international organizations, research centers, and professional organizations.
With about 100 sets of principles published so far, it is easy to get lost in these separate but similar documents. We know, because we found ourselves struggling to keep up with the trend.

We decided to find a solution both for our own sake and for others in the field.
And we created the **Dynamics of AI Principles**.

Dynamics of AI Principles



Created by:

K. Yasemin Usta
Dima Al-Qutub
Oguz Kartoz
&
Cansu Canca

(February 2020)

aiethicslab.com/big-picture/

Using the Dynamics of AI Principles

With this interactive toolbox, you can;

I. use *the Map* to **sort, locate**, and **visualize** AI principles by
   a. country and region,
   b. time of publication,
   c. types of publishing organizations;
II. **search** documents or see the full list and find their summaries,
III. **compare** documents and their **key points**,
IV. visualize and compare the **distribution** of core principles, and
V. use *the Box* to **systematize principles** and **evaluate technologies**.

AI ETHICS LAB

What are AI Principles for?

Over 100 sets of AI principles… But what do they mean in practice?

What are principles for and how can we operationalize them for building ethical AI technologies?

Principles help us **recognize** and **keep in mind the ethical considerations** we must take into account when we make decisions.

They provide a valuable tool for **detecting** and **conceptualizing** ethical concerns that a technology poses.

What are AI Principles not for?

Principles are **not** coherent and complete systems for decision-making.

They are loosely based on moral and political theories. Each theory provides a coherent and structured value system and a decision-making tool. Principles capture the main ideas of these theories. But they also capture the conflicts between these theories without offering a resolution.

Principles can help us think, they **cannot** systematically guide us.

Categorizing the Principles

So what is the structure behind this abundance of principles?

We can categorize all principles into **3 core** principles:

respect autonomy, avoid harm & do good, and ensure justice.

These principles also form the *Principlism* framework—the most dominant principle-based framework in applied ethics for over 40 years.

3 Core Principles: Intrinsic Values

Pro tip: Use the Dynamics page to unfold the core principles.

No Harm

Autonomy

Benefit

Justice

These 3 core principles are different than others because they are *intrinsically* valuable whereas many other principles are mere explications of these four.

There is no hierarchy between the 3 core principles and none of them can be sacrificed for another. In other words, if and when these principles conflict, you are faced with a real ethical dilemma.

AI ETHICS LAB

Instrumental Principles

Other principles—such as *transparency, explainability, privacy, and accountability*— are **instrumental**.

Autonomy

- Power to decide (whether to decide)
- Human control
- Human oversight
- Transparency (to understand)
- Openness (to understand)
- Explainability
- Explicability
- Liberty
- Freedom
- Fundamental rights
- Personal privacy
- Privacy protection
- Fundamental rights
- Human values

No Harm

- Control risks
- Safety
- Security
- Capability caution
- Data protection
- Privacy (to avoid harm)
- Explicability
- Transparency (to avoid harm)
- Reproducibility
- Accuracy
- Reliability
- Responsible deployment
- Prevent arms race

Benefit

- Promoting well-being
- Benefit society
- Generating net benefits
- Sustaining the planet
- Impact
- Efficacy
- Explicability
- Scientific excellence
- User-centered design (for user benefit)
- People-first approach

Justice

- Fairness
- Fundamental rights
- Equality
- Non-discrimination
- Avoiding bias
- Inclusivity
- Diversity
- Data neutrality
- Representative data
- Shared benefit / prosperity
- Social & economic impacts
- Avoid disparity
- Mitigating social dislocation
- Preserving solidarity
- Accessibility
- Explicability
- Transparency (for accountability)
- Openness (for accountability)
- Accountability
- Auditability
- Liability
- Inclusive
- Judicial transparency
- Open governance
- Regulatory & legal compliance

What does this mean?

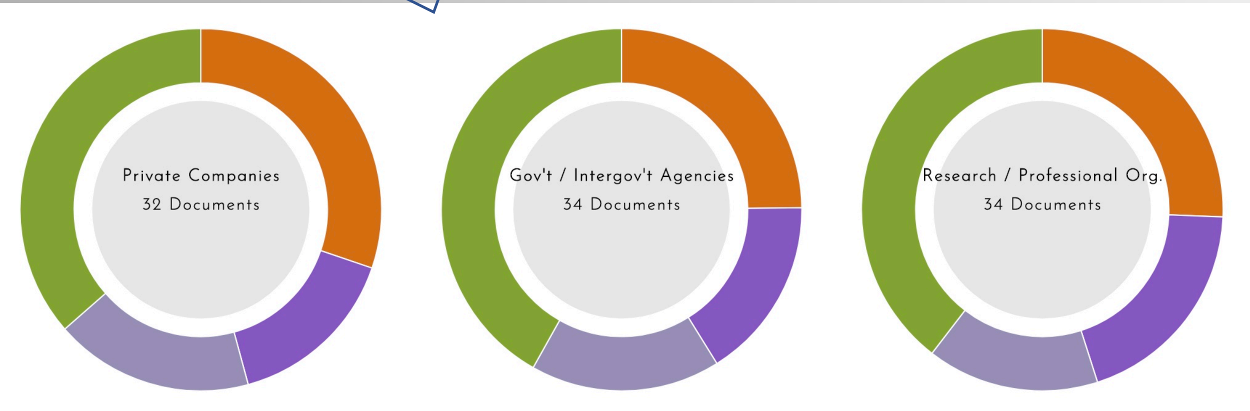
It means that these principles help **protect** and **promote** the **core** principles.

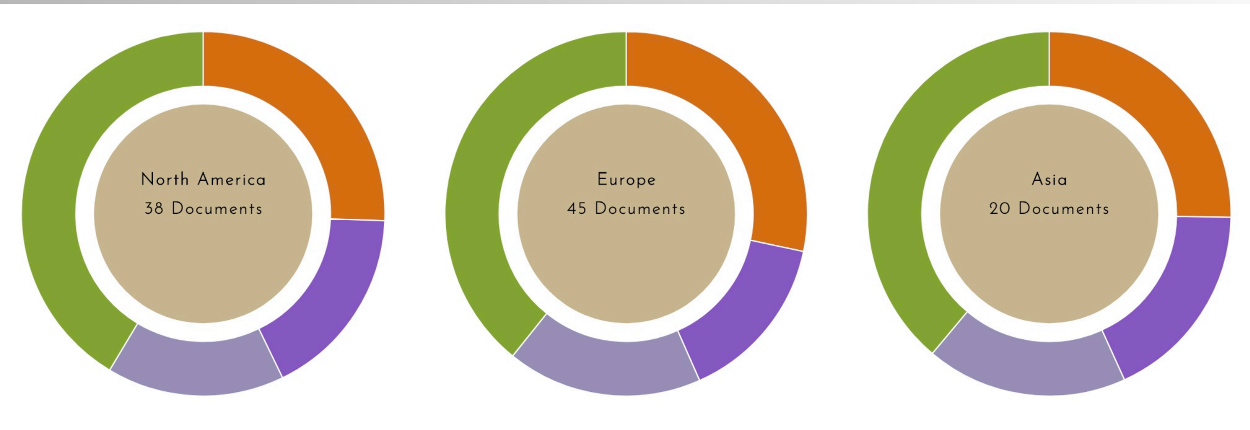
It also means that if these instrumental principles conflict, it is <u>not</u> a dilemma because instrumental principles are interchangeable.

Categorizing the Principles

Pro tip: Use the Dynamics page to check each principle's ratio and their combined values.

Once categorized into core principles, we find that similar weights are given to each principle across organizations and regions:





Of all principles;

- ~25–30% are *autonomy*-focused,
- ~32–34% are focused on avoidance of *harm* & increasing *benefits*,
- ~36–41% are *justice*-focused.

Operationalizing the Principles

Let's put these principles into use and see how the core & instrumental principles play out in practice.

We'll go over a use-case (Google AI Principles x Google Duplex) and demonstrate 3 steps:

Step 1: Organize *instrumental* principles into the 3 *core* principles

Step 2: Lay out ethical concerns of the case using the *organization's instrumental* principles

Step 3: Check if you need to add more instrumental principles to uphold the 3 core principles

CASE: Google x Duplex

GOOGLE AI PRINCIPLES

1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.
7. Be made available for uses that accord with these principles.

Read the full document:
https://ai.google/principles/

GOOGLE DUPLEX

Google announced “a new technology for conducting natural conversations to carry out ‘real world’ tasks over the phone. The technology is directed towards completing specific tasks, such as scheduling certain types of appointments. For such tasks, the system makes the conversational experience as natural as possible, allowing people to speak normally, like they would to another person, without having to adapt to a machine.”

Read the full document:
https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

AI ETHICS LAB

CASE: Google x Duplex

Step **1**: Organize *instrumental* principles into 3 core principles

Autonomy
Harm–Benefit
Justice

GOOGLE AI PRINCIPLES

1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.
7. Be made available for uses that accord with these principles.

Read the full document:
https://ai.google/principles/

AI ETHICS LAB

CASE: Google x Duplex

Step **1**: Organize *instrumental* principles into 3 core principles

GOOGLE AI PRINCIPLES

Autonomy

- Incorporate **privacy** design principles.

Harm–benefit

- Be built and tested for **safety**.
- Be **socially beneficial**.
- Uphold high standards of **scientific excellence**.

Justice

- Avoid creating or reinforcing **unfair bias**.
- Be **accountable** to people.
- Be made available for uses that accord with these **principles**.

AI ETHICS LAB

CASE: Google x Duplex

Step **2**: Lay out ethical concerns using the *organization's instrumental* principles

GOOGLE AI PRINCIPLES

Autonomy

- Incorporate **privacy** design principles.

Harm–benefit

- Be built and tested for **safety**.
- Be **socially beneficial**.
- Uphold high standards of **scientific excellence**.

Justice

- Avoid creating or reinforcing **unfair bias**.
- Be **accountable** to people.
- Be made available for uses that accord with these **principles**.

GOOGLE DUPLEX

Autonomy

- privacy: ensure caller & receiver privacy

Harm–benefit

- safety: address misuse (e.g., scam, impersonation, fake information)
- social benefit: help those who have speech problems, reduce cost for (small & large) businesses
- scientific excellence: ensure testing & development of the system

Justice

- non-discrimination: understand & imitate diverse types of speech
- accountability: ensure mechanism for accountability
- ethical process: test for ethical research, development, design, and deployment of the system

AI ETHICS LAB

CASE: Google x Duplex

Step **3**: Check if you need other instrumental principles to uphold 3 core principles?

GOOGLE AI PRINCIPLES

Autonomy
- Incorporate **privacy** design principles.

Harm–benefit
- Be built and tested for **safety**.
- Be **socially beneficial**.
- Uphold high standards of **scientific excellence**.

Justice
- Avoid creating or reinforcing **unfair bias**.
- Be **accountable** to people.
- Be made available for uses that accord with these **principles**.

THE BOX – by AI Ethics Lab

		None	Minimum	Medium	Maximum	
Autonomy	Human control / oversight					N / A
	Transparency					N / A
	Explainability					N / A
	Information					N / A
	Agency					N / A
	Consent					N / A
	Privacy					N / A
Harm-Benefit	Accuracy / Reliability					N / A
	Security					N / A
	Safety					N / A
	Well-being					N / A
	Impact					N / A
	Efficiency					N / A
Justice	Distribution of burden & benefit					N / A
	Equality / Non-discrimination					N / A
	Protecting the vulnerable					N / A
	Accountability					N / A
	Contestability					N / A

Use *the Box* tool to check for other relevant principles. You can also rate the technology for how it scores on each principle.

Find *the Box* on the Dynamics page. To learn more about *the Box*, click here.

CASE: Google x Duplex

Step **3**: Check if you need other instrumental principles to uphold 3 core principles?

GOOGLE AI PRINCIPLES

Autonomy

- Incorporate **privacy** design principles.

Harm–benefit

- Be built and tested for **safety**.
- Be **socially beneficial**.
- Uphold high standards of **scientific excellence**.

Justice

- Avoid creating or reinforcing **unfair bias**.
- Be **accountable** to people.
- Be made available for uses that accord with these **principles**.

GOOGLE DUPLEX

Autonomy

- privacy: ensure caller & receiver privacy
- **transparency**: clarify in engaging with AI vs. human
- **agency**: avoid deception to ensure user understanding & choice

Harm–benefit

- safety: address misuse (e.g., scam, impersonation, fake information)
- social benefit: help those who have speech problems, reduce cost for (small & large) businesses
- scientific excellence: ensure testing & development of the system
- **avoid societal harm**: avoid erosion of social trust & social relations

Justice

- non-discrimination: understand & imitate diverse types of speech
- accountability: ensure mechanism for accountability
- ethical process: test for ethical research, development, design, and deployment of the system

AI ETHICS LAB

CASE: Google x Duplex

Must address **(at the minimum!)**:

- Privacy, **transparency, agency**
- Safety, social benefits, scientific excellence, **societal harm**
- Non-discrimination, accountability, ethical process

(Keep in mind, these points help us think. They do not provide conclusive guidance.)

Duplex (announced in May 2018) caused ethical outrage because Google (after announcing its AI Principles in March 2018) completely neglected the deception that Duplex displays and the risks that this deception poses to social relations as well as individual and collective decision-making.

What makes the Duplex case a good illustration is that

(1) the ethical neglect was due to limited understanding of core principles &

(2) this ethical error was completely avoidable.

AI ETHICS LAB

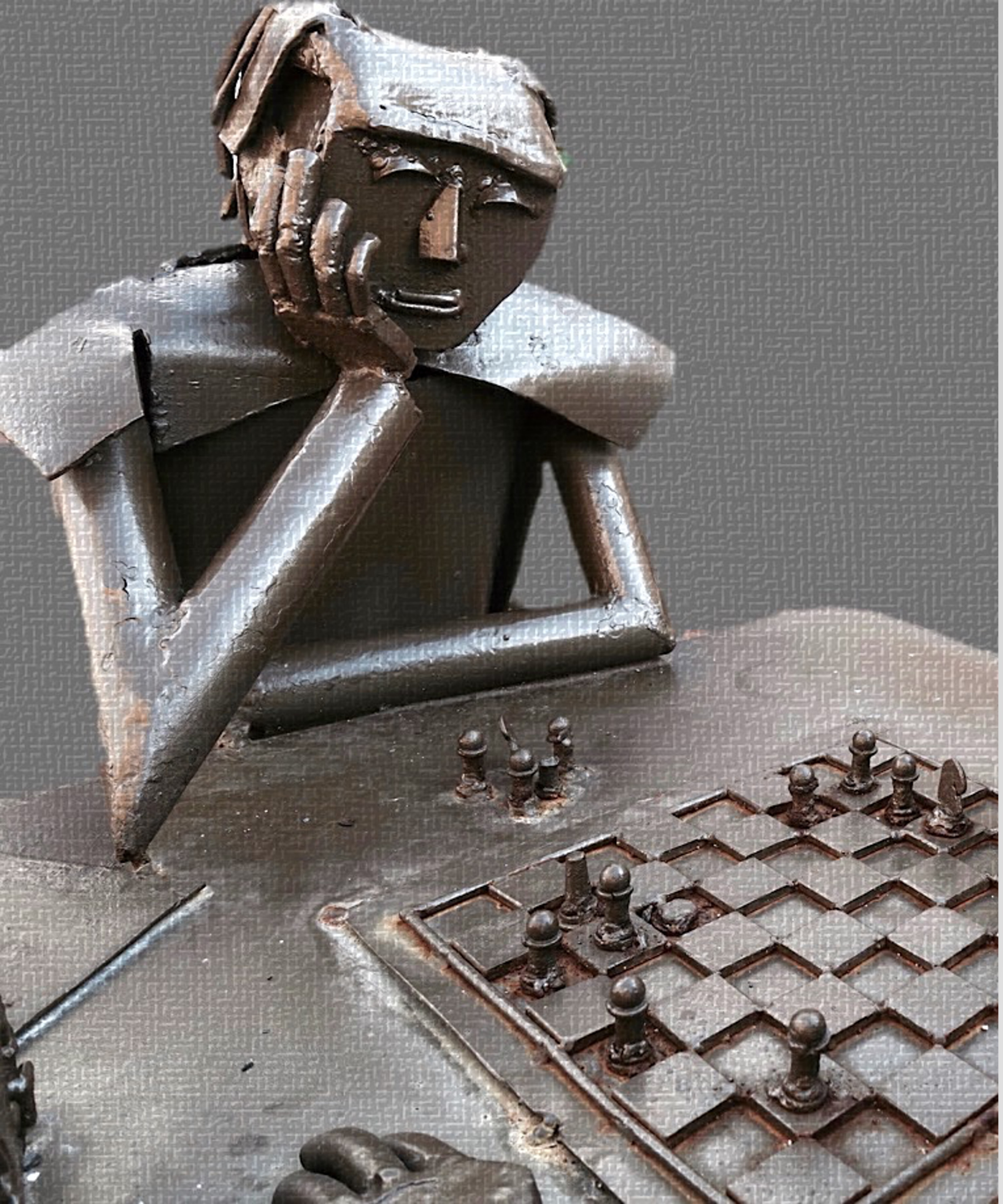
Moving Beyond the List

What's the use of defining your AI Principles?

Once categorized into core & instrumental principles, AI Principles will help;

- ✓ set the Ethics Strategy of your institution / company,
- ✓ frame Ethics Trainings to guide practitioners in simple cases,
- ✓ guide decision-makers in complex cases after the Ethics Analysis by ethics experts has laid out the ethically justified options.

As with any other tool, AI Principles are useful only if utilized correctly!



Want to learn more?

Visit **Dynamics of AI Principles**:

aiethicslab.com/big-picture/

Visit **AI Ethics Lab**:

aiethicslab.com

Contact us:

contact@aiethicslab.com

AI ETHICS LAB