A Guide to the Dynamics of AI Principles:

Making Sense of AI Principles & How to Use Them

AI ETHICS LAB

February 2020

A Guide to the Dynamics of AI Principles

This guide will help you:

- learn more about our tool, **Dynamics of AI Principles**,
- understand the structure behind AI principles,
- systematize AI principles for your own use, and
- operationalize AI principles for your company.

Let's get started! 🏁

AI ETHICS LAB

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.

AI ETHICS LAB

What is the Dynamics of AI Principles?

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

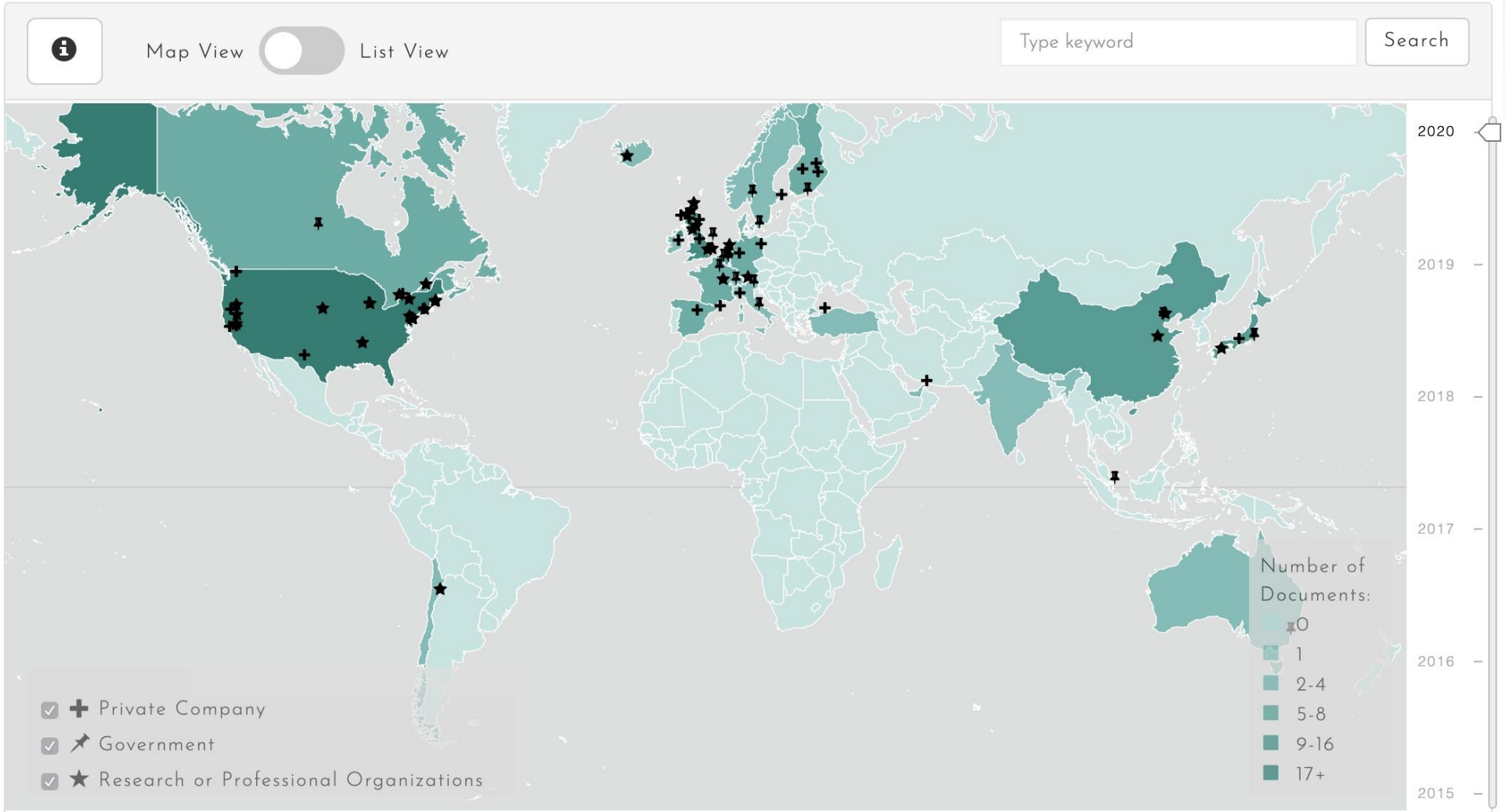
AI Principles are published by private companies, governmental agencies, international organizations, research centers, and professional organizations.
With about 100 sets of principles published as of today, it is easy to get lost in these separate but similar documents. We know, because we found ourselves 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**.

AI ETHICS LAB

Dynamics of AI Principles



Compare AI principles

Number of documents: 3

Name: Choose Document

Name: Choose Document

Name: Choose Document

Created by:

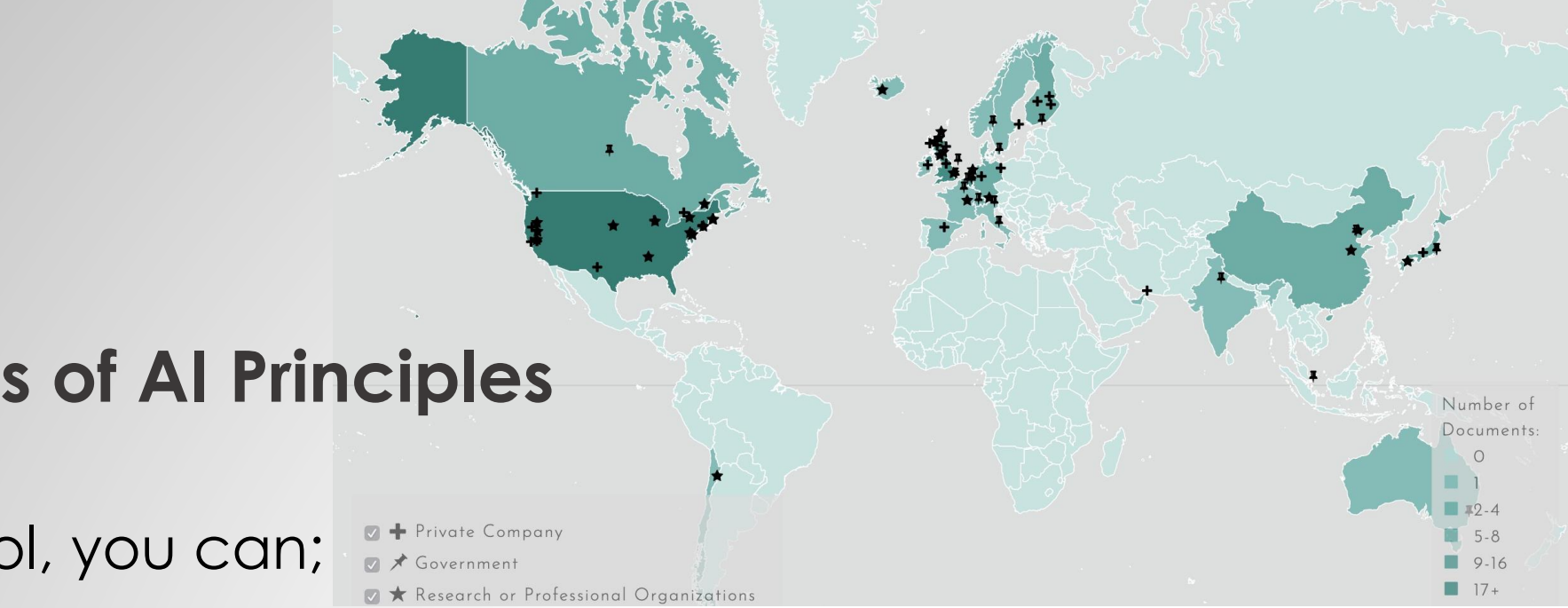
K. Yasemin Usta
Dima Al-Qutub
Oguz Kartoz
&
Cansu Canca

(February 2020)

aiethicslab.com/big-picture/

AI ETHICS LAB

Using the **Dynamics of AI Principles**



With this interactive tool, you can;

I. **sort, locate,** and **visualize** AI principles by
   a. country and region,
   b. time of publication,
   c. types of publishing organizations;
II. **search** for a particular document and get its summary,
III. see the full list **alphabetically**,
IV. **compare documents and their summaries!**

AI ETHICS LAB

What are AI Principles for?

We have all these 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 **remember the ethical considerations** we must take into account when we make decisions *(see: our use-case)*.

Some principles are intrinsically valuable and some are only instrumental.

AI ETHICS LAB

What are AI Principles **not** for?

Principles are **not** coherent and structured systems.

They are loosely based on moral and political theories. Each of these theories provide coherent and structured value systems. Principles capture their main ideas without putting them into an applicable framework.

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

AI ETHICS LAB

Categorizing the Principles

So what is the structure behind this abundance of principles?

We can categorize all principles into **4 core** principles:

These principles also form the *Principlism* framework—the most dominant principle-based framework in applied ethics for the last 50 years.

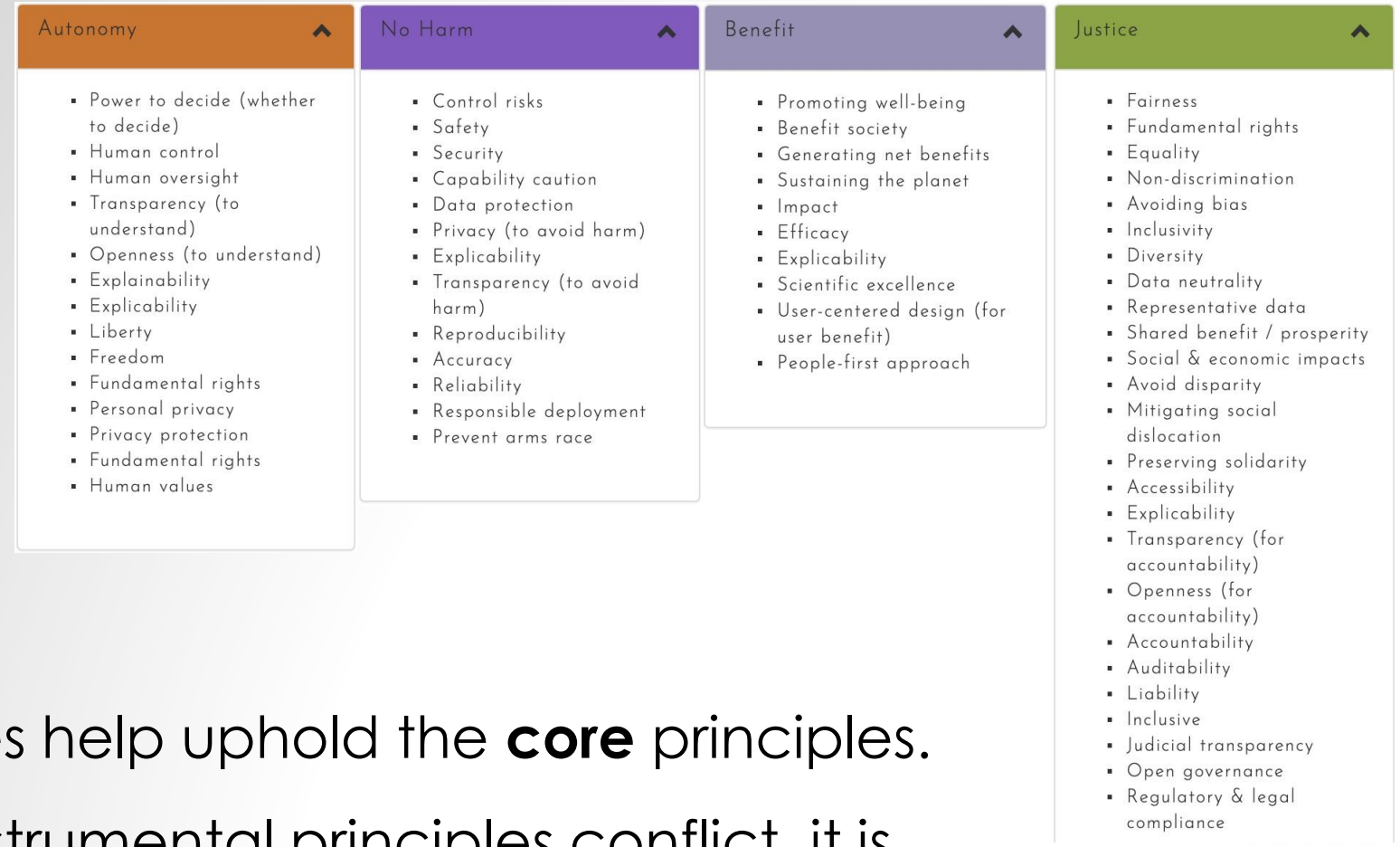
4 Core Principles: Intrinsic Values

These 4 core principles are different than others because they are *intrinsically* valuable and many other "principles" are mere explications of these four.

There is no hierarchy between the 4 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.

Instrumental Principles

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

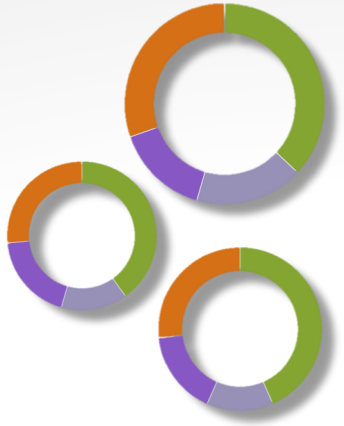


What does this mean?

It means that these principles help uphold the **core** principles.

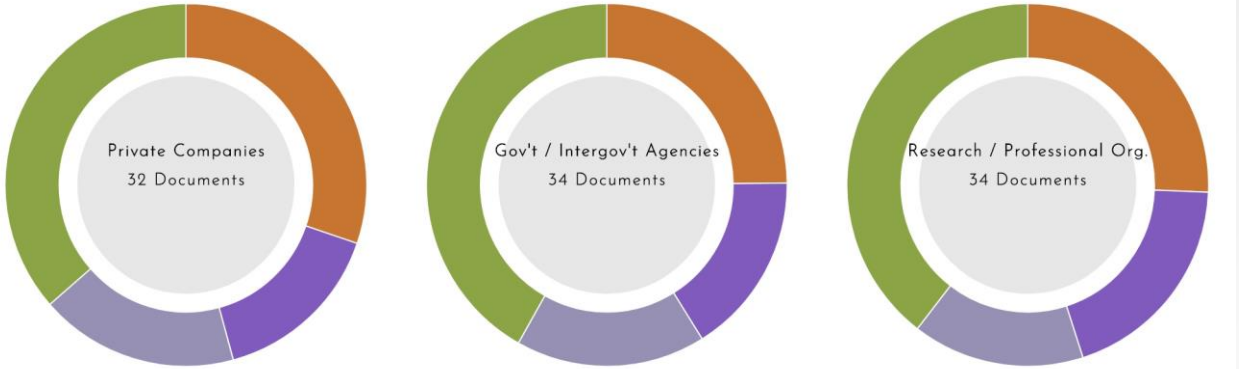
It also means that if these instrumental principles conflict, it is *not* a dilemma.

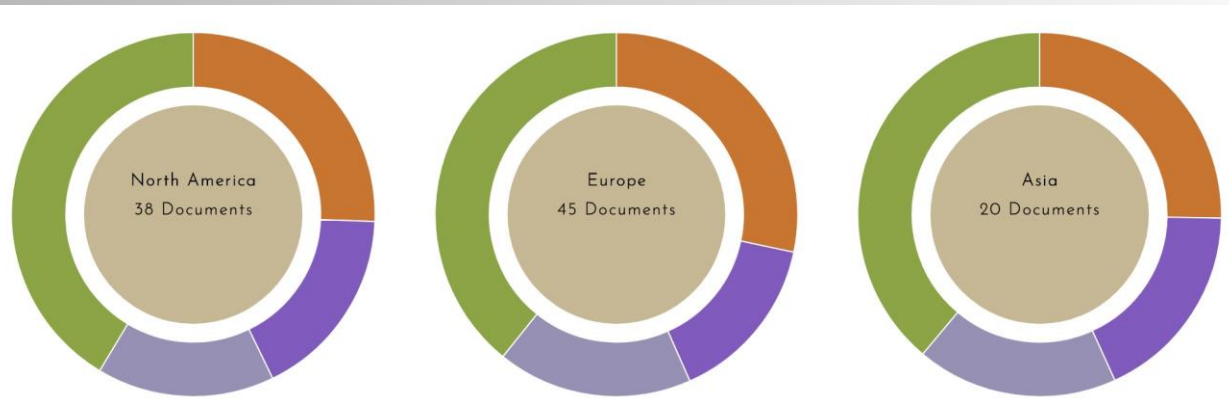
Categorizing the Principles



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

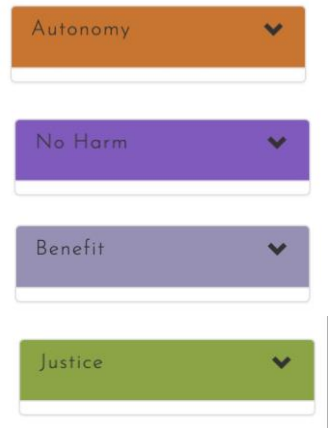
Divided by the type of organization that published them or the region where they were published, a consistent picture emerges: Of all principles;





- ~25–30% are *autonomy*-focused,
- ~33–36% are focused on avoidance of *harm* & increasing *benefits*,
- ~36–42% are *justice*-focused

regardless of how we divide the existing list.



AI ETHICS LAB

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 Duplex) and demonstrate 3 steps:

Step 1: Organize *instrumental* principles into the 4 core principles

Step 2: Lay out ethical concerns of the case using the *stated instrumental* principles

Step 3: Check if you need to add more instrumental principles to uphold the 4 core principles

AI ETHICS LAB

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."

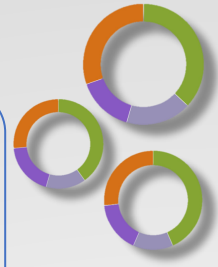
Read the full document:
https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

AI ETHICS LAB

CASE: Google x Duplex

Step **1**: Organize *instrumental* principles into the 4 core principles

Autonomy
Avoid harm
Maximize benefits
Justice



GOOGLE AI PRINCIPLES

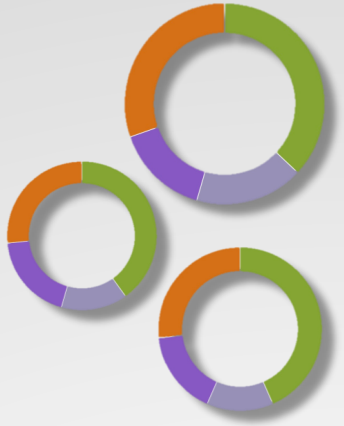
1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.
7. Be made available for uses that accord with these principles.

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

AI ETHICS LAB

CASE: Google x Duplex

Step **1**: Organize *instrumental* principles into 4 core principles



GOOGLE AI PRINCIPLES

Autonomy
- Incorporate privacy design principles.

Avoid harm
- Be built and tested for safety.

Maximize benefits
- 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 of the case using the *stated instrumental* principles

GOOGLE AI PRINCIPLES

Autonomy
- Incorporate privacy design principles.

Avoid harm
- Be built and tested for safety.

Maximize benefits
- 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

Avoid harm
- safety: address misuse (e.g., scam, impersonation, fake information)

Maximize benefits
- social benefit: help those who have speech problems, reduce cost for (small & large) businesses
- scientific excellence: enable testing & development of the system

Justice
- bias: understand & imitate diverse types of speech
- accountability: ensure mechanism for accountability
- principles: test for ethical research, design, development & deployment of the system

AI ETHICS LAB

CASE: Google x Duplex

Step **3**: Do you need to add more instrumental principles to uphold 4 core principles?

GOOGLE AI PRINCIPLES

Autonomy
- Incorporate privacy design principles.

Avoid harm
- Be built and tested for safety.

Maximize benefits
- 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
- agency: avoid deception to ensure user understanding & choice
- transparency: clarify in engaging with AI vs. human

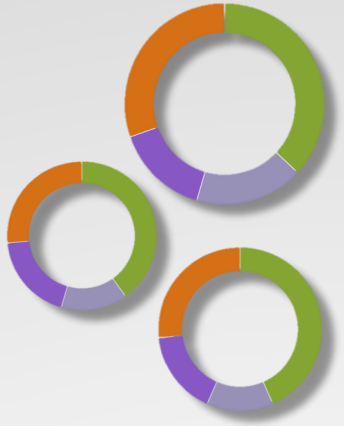
Avoid harm
- safety: address misuse (e.g., scam, impersonation, fake information)
- societal well-being: avoid erosion of social trust & social relations

Maximize benefits
- social benefit: help those who have speech problems, reduce cost for (small & large) businesses
- scientific excellence: enable testing & development of the system

Justice
- bias: understand & imitate diverse types of speech
- accountability: ensure mechanism for accountability
- principles: test for ethical research, design, development & deployment of the system

AI ETHICS LAB

CASE: Google x Duplex



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

- Privacy, **agency, transparency**
- Safety, avoid misuse, **societal well-being**
- Social benefits, scientific excellence
- Avoid bias, 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.

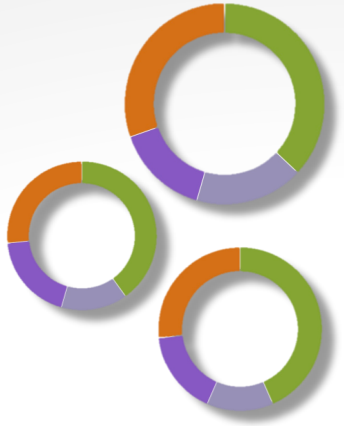
What makes the Duplex case a good illustration is that

(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



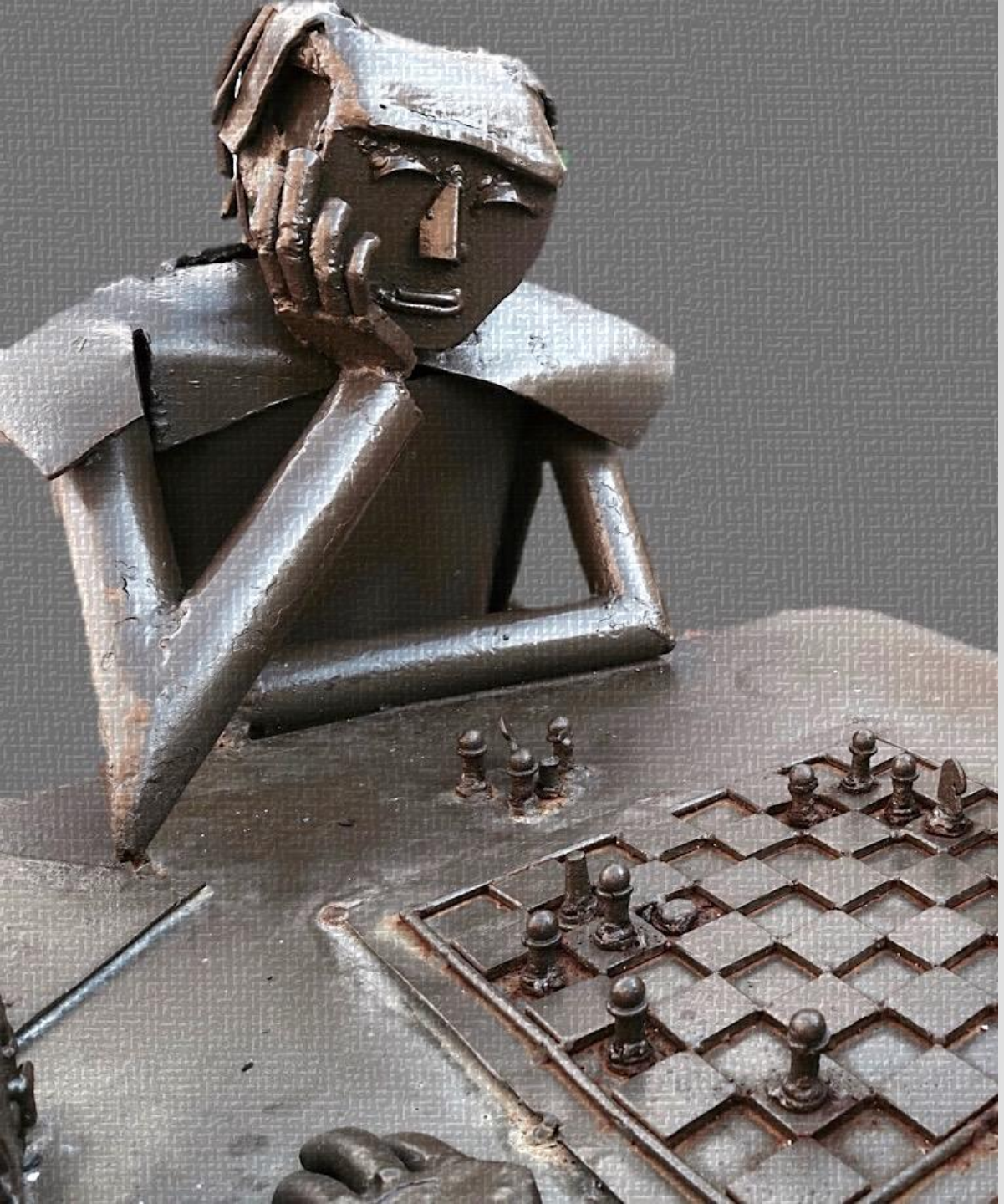
What's the use of defining your AI Principles?

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 hard cases after the Ethics Analysis has laid out the ethically justified options.

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

AI ETHICS LAB



Want to learn more?

Visit **Dynamics of AI Principles**:

aiethicslab.com/big-picture/

Visit **AI Ethics Lab**:

aiethicslab.com

Contact us:

contact@aiethicslab.com