How to Use The Box

A Tool for Operationalizing AI Principles

AI ETHICS LAB

June 2020

What is the Box?

The Box is a tool from our *Dynamics of AI Principles* toolkit.

You can use it to think through the ethical implications of the technologies that you are evaluating or creating.

The Box is a simplified tool that

- lists important ethical principles and concerns,
- puts instrumental ethical principles in relation to core principles,
- helps visualize ethical strengths & weaknesses of technologies, and
- enables visual comparison of technologies.

AI ETHICS LAB

Structure of the Box

The Box presents core ethical principles *(i.e., respecting autonomy, avoiding harm & doing good, ensuring justice)* and the instrumental principles that primarily link to them.

CORE PRINCIPLES	INSTRUMENTAL PRINCIPLES
Autonomy	human control (oversight), transparency, explainability, information, agency, consent, privacy
Harm-Benefit	accuracy, reliability, security, safety, well-being, impact, efficiency
Justice	distribution of burdens & benefits, equality, non-discrimination, protecting the vulnerable, accountability, contestability

Using the Box

Click here for the printable version.

Click here to unfold the core principles.

Click here to open this manual at any time.

If a principle does not apply, use the N/A button to disable the slider.

Use the sliders to rate how strongly each principle applies to the technology you are evaluating.

Consult the questions on page 1 for the details of the principles.

QUESTIONS TO ASK

Autonomy	human control / oversight	*Does the system allow for human oversight and control?* *Does the system provide the necessary information for meaningful oversight and control?*
	transparency	*Are the system's abilities and limitations clear to the users and those who are subject to the system?* *Is the system transparent in how it affects individual decision-making?*
	explainability	*Is it explainable how the system reaches its decisions and outcomes?* *Can humans understand how the system produces its results?*
	information	*Does the system provide individuals with accurate and relevant information regarding the system?* *Does the system enable individuals' access to accurate and relevant information for decision-making?*
	agency	*Does the system enable individuals pursue their goals or help their pursuit?*
	consent	*Is the system designed to ensure rational, voluntary, and informed consent of individuals when they use its functions?*
	privacy	*Does the system allow individuals control their privacy? Does the system protect individual privacy?*
Harm-Benefit	accuracy / reliability	*Is the system accurate and reliable? Is the system designed to be measured for accuracy and reliability?*
	security	*Is the system secure and designed against security attacks?*
	safety	*Is the system safe for users and those who are subject to the system?* *Is the system designed to minimize risks of harm from the system or any other risks of harm?*
	well-being	*Does the system promote and protect individual and societal well-being?* *Does the system maximize well-being and benefits in society?*
	impact	*Does the system have a significant beneficial impact? Is the system designed to be measured for impact?*
	efficiency	*Is the system efficient in achieving the set goals? Is the system designed to be measured for efficiency?*
Justice	distribution of burden & benefit	*Is the system designed and developed to ensure (1) not putting disproportionate burden on a group and (2) not concentrating benefits on another group*
	equality / non-discrimination	*Is the system developed with an understanding of social equality and aim to ensure it?* *Is the system designed to reduce or eliminate unfair bias?*
	protecting the vulnerable	*Does the system protect vulnerable individuals and the worst-off from harm and aim to benefit them?*
	accountability	*Is the system designed for accountability and auditability? Is the system designed for traceability?*
	contestability	*Does the system have a mechanism for appeals?*

AI ETHICS LAB

CASE: Designing a Hiring Algorithm

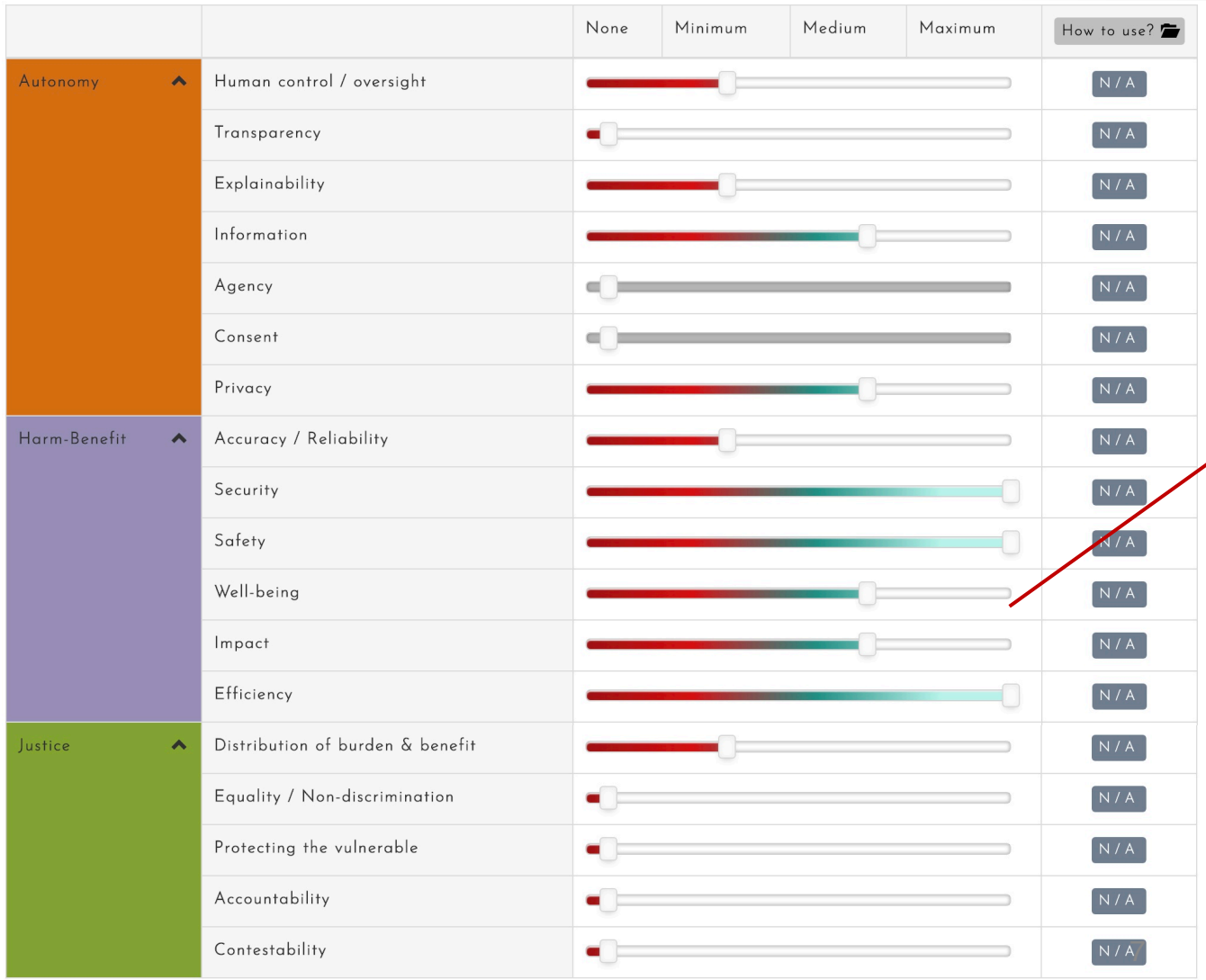
Your team is building a hiring algorithm for a technology firm using a machine learning model. The algorithm will be used by the HR to do the first round of shortlisting of candidates. Only those CVs that pass a threshold score given by the algorithm will be reviewed by the HR.

The firm is large and over 20 years old. You decide to use historical hiring data.

You know that the leadership positions of the firm has always been heavily male dominated. Therefore, you make sure to mask the gender so that the algorithm does not use the gender of the applicant as a parameter.

AI ETHICS LAB

CASE: Designing a Hiring Algorithm



Low scores on transparency, explainability, & human control:

The system is a "black box" and its function and limitations are unclear to the applicants and to the HR.

Average scores on well-being & impact:

The system might be overall beneficial and impactful.

Low scores on justice:

The system uses historical data that is most likely to be biased against female applicants. Masking the gender is unlikely to be enough to solve this problem.

The system also does not have mechanisms for appeal and accountability.

AI ETHICS LAB

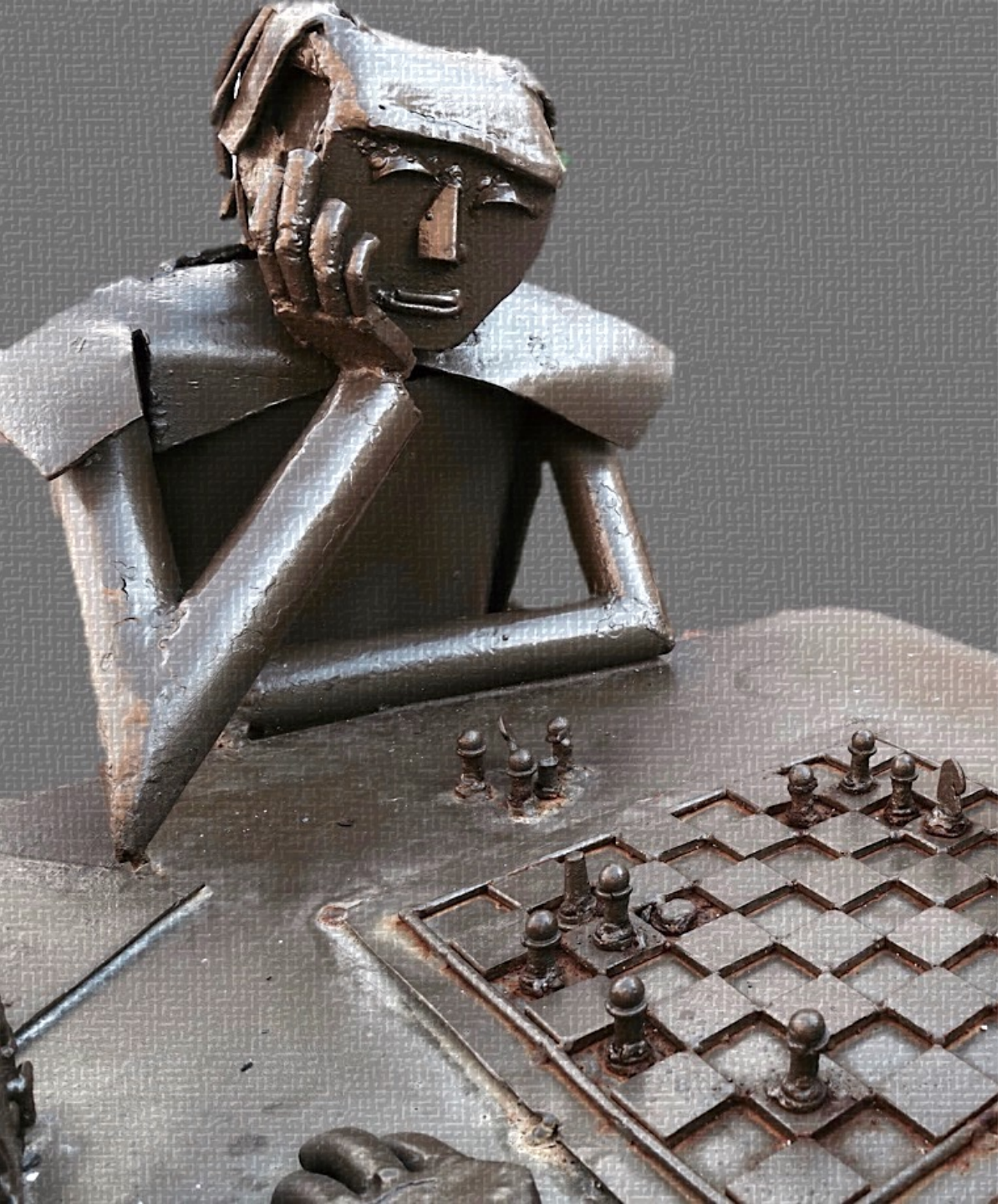
The Box: Benefits & Limitations

The Box helps operationalize AI ethics principles. This is only the *start* of an ethics analysis. Once the strengths and weaknesses of a technology is revealed, then we need to determine the complexity of the ethical issue at hand and plan solutions accordingly.

Depending on the types of ethical problems that arise, solving them might require technical, design, and/or ethics experts to come in.

For more on principles, see the *Dynamics Guide* (pages 6-11) and the article *"Operationalizing AI Ethics Principles"*.

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

AI ETHICS LAB