TOWARDS TRANSPARENT AND ACCOUNTABLE AI IN PUBLIC SERVICE

DAI

Shaping a more livable world.

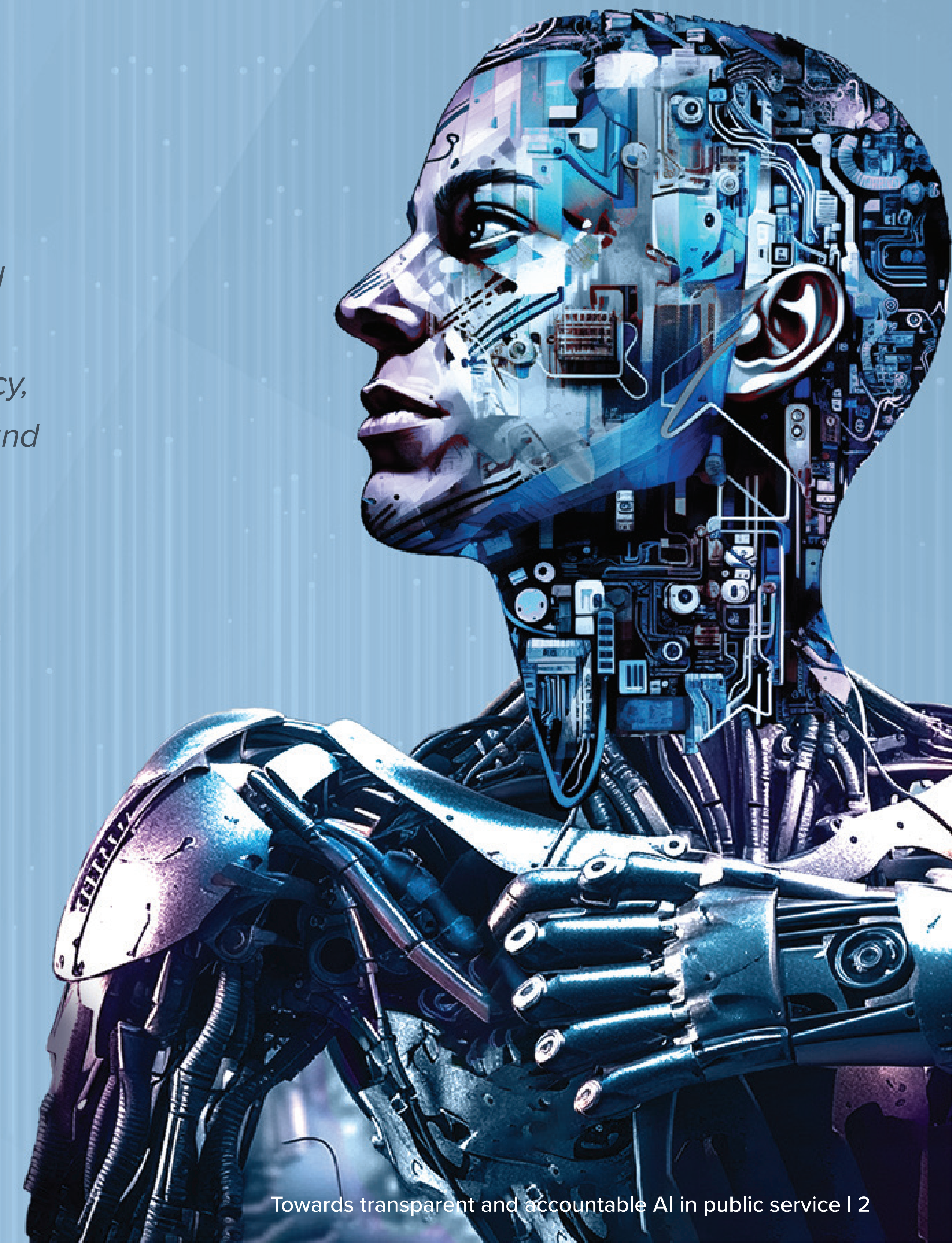
ARTIFICIAL INTELLIGENCE

AI impacts everyone, performing complex, risky, or monotonous tasks, aiding doctors and lawyers, and automating public services. However, algorithmic systems may violate human rights, lack transparency, and reinforce discrimination due to biases in data and programming, leading to errors in novel situations.

Governments worldwide are using AI algorithms to automate or support decision-making in public services.

Algorithms are used in urban planning, social care, welfare, unemployment fraud detection, and criminal justice.

The use of AI algorithms is often seen as a way to improve efficiency and reduce costs of public services.

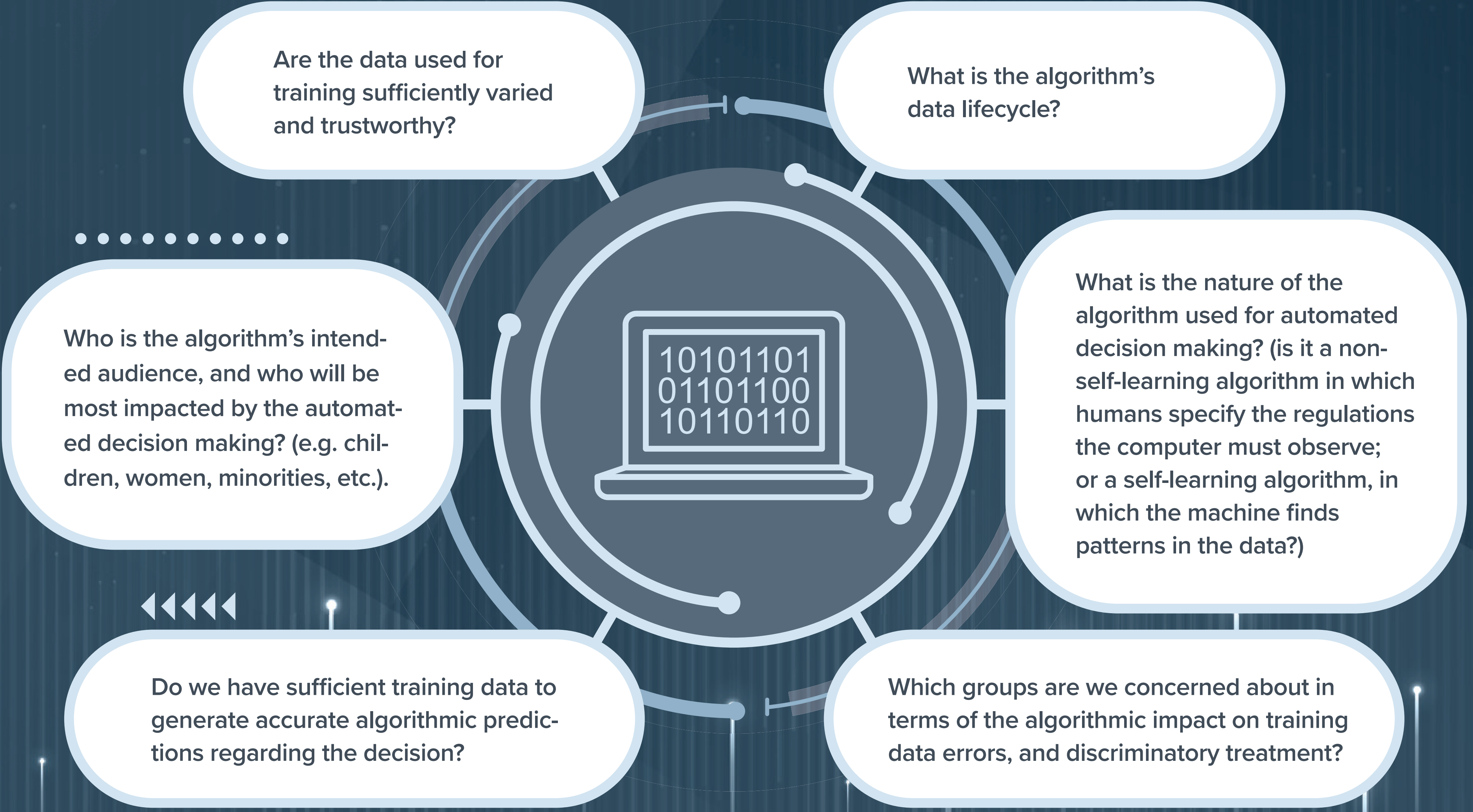


MAPPING, CONCEPTUALIZATION, AND INITIAL ANALYSIS

Questions to Consider

DAI

WHAT WILL THE AUTOMATED DECISION DO?



Are the data used for training sufficiently varied and trustworthy?

What is the algorithm's data lifecycle?

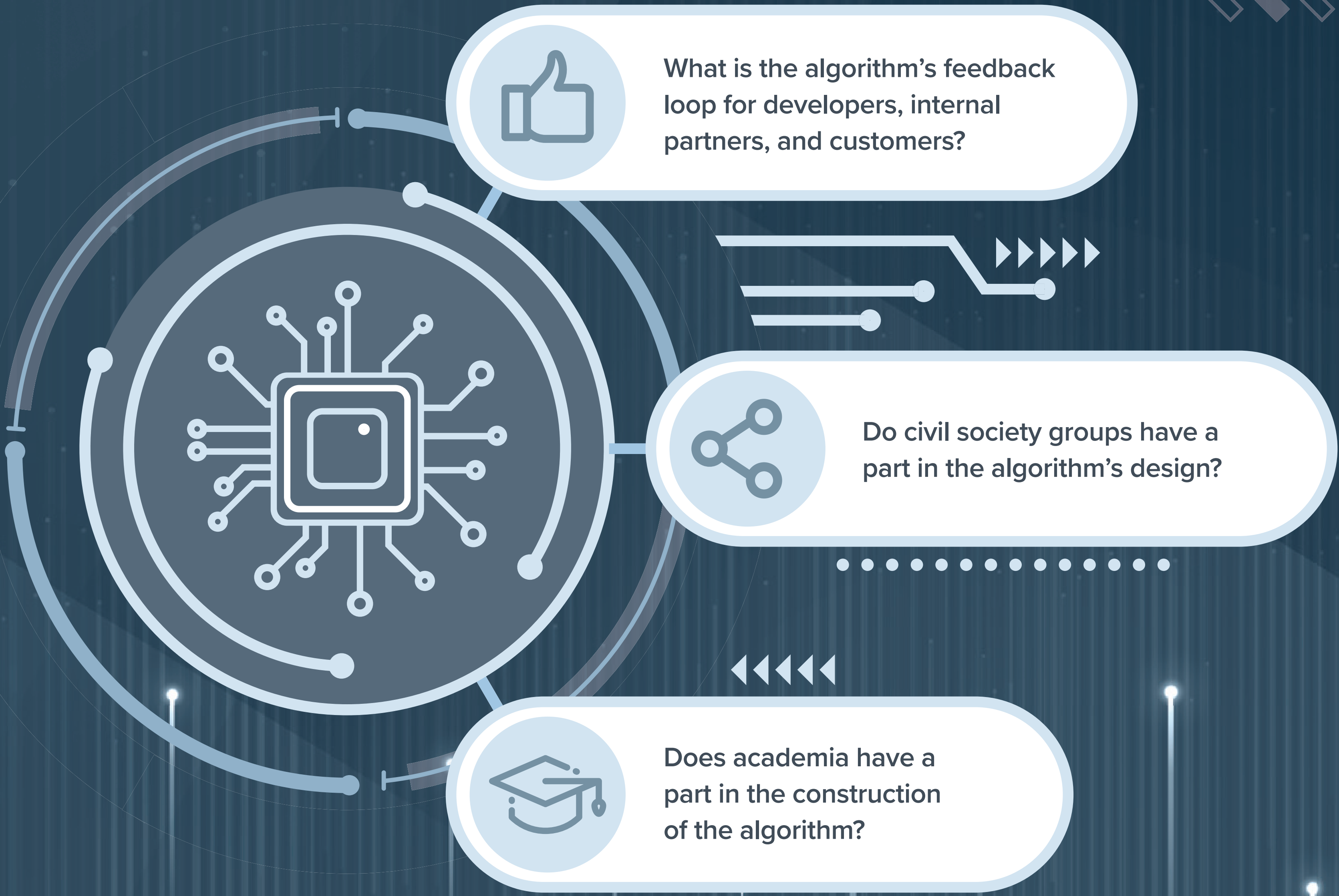
Who is the algorithm's intended audience, and who will be most impacted by the automated decision making? (e.g. children, women, minorities, etc.).

What is the nature of the algorithm used for automated decision making? (is it a non-self-learning algorithm in which humans specify the regulations the computer must observe; or a self-learning algorithm, in which the machine finds patterns in the data?)

Do we have sufficient training data to generate accurate algorithmic predictions regarding the decision?

Which groups are we concerned about in terms of the algorithmic impact on training data errors, and discriminatory treatment?

HOW ARE ADDITIONAL STAKEHOLDERS ENGAGED?



What is the algorithm’s feedback loop for developers, internal partners, and customers?

Do civil society groups have a part in the algorithm’s design?

Does academia have a part in the construction of the algorithm?

HAS DIVERSITY BEEN TAKEN INTO ACCOUNT IN THE DESIGN AND IMPLEMENTATION?

Will the algorithm affect particular cultural groups and behave differently in cultural contexts?

Is the design team sufficiently diverse to capture cultural subtleties and foresee the algorithm’s applicability in various cultural contexts?

If not, what measures do we have in place to make these scenarios more prominent and comprehensible to designers?

Considering the objective of the algorithm, are the training data sufficiently diverse?

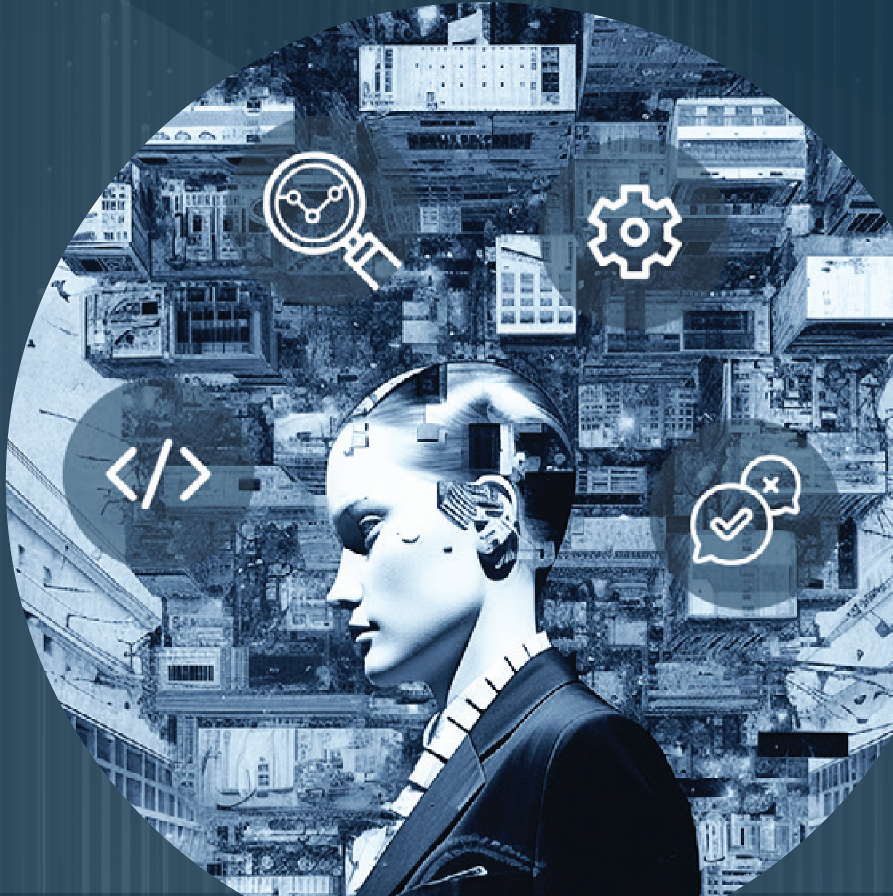
Are there statutory guidelines that public sector organizations should check to ensure that the application of the algorithm is legal and ethical?

WHAT IS THE LEGAL BASIS FOR AUTOMATED DECISION MAKING?

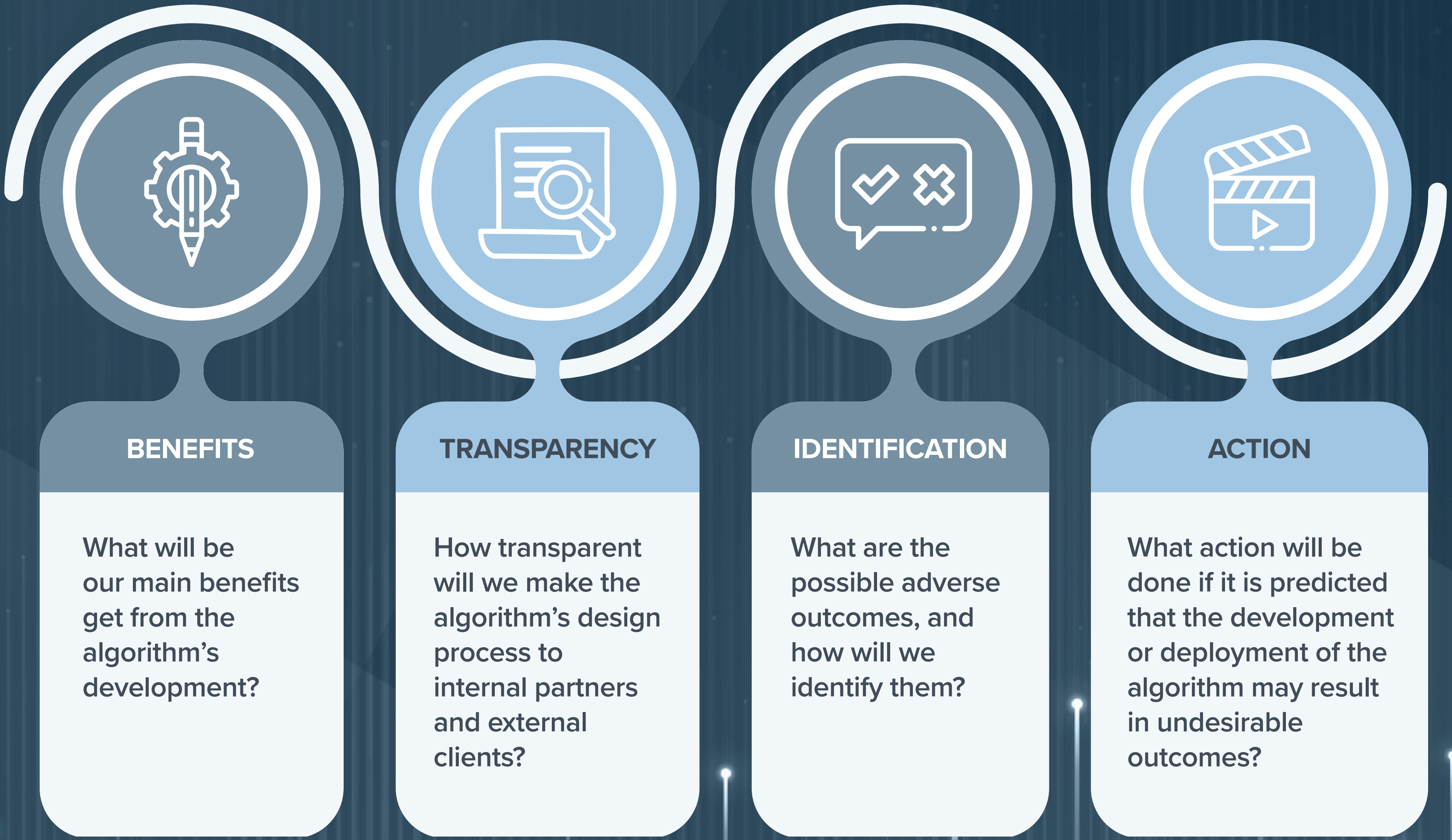
If an algorithm is expected to affect human rights, there must be a legal basis for its use.

WHAT ARE THE OBJECTIVES OF THE AUTOMATED DECISION MAKING PROCESS?

Why is the algorithm needed and what outcomes is it intended to enable?



WHAT ARE THE INCENTIVES FOR AUTOMATED DECISION MAKING?



BENEFITS

What will be our main benefits get from the algorithm's development?

TRANSPARENCY

How transparent will we make the algorithm's design process to internal partners and external clients?

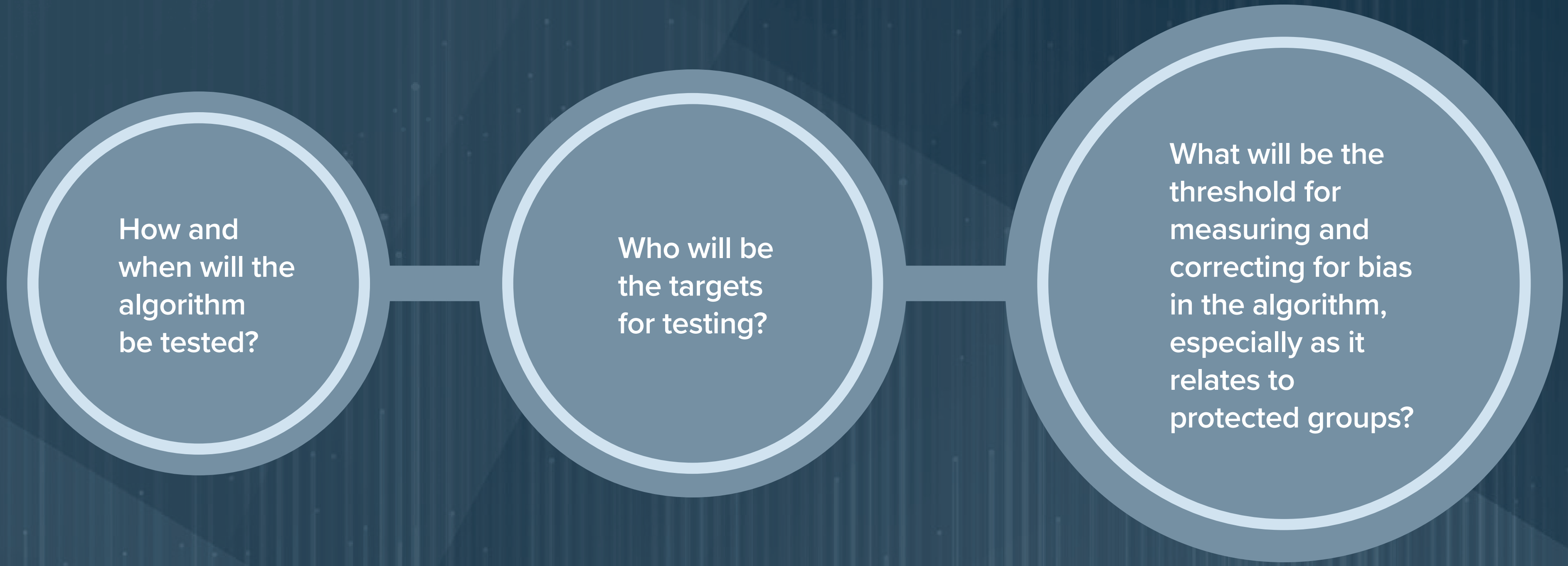
IDENTIFICATION

What are the possible adverse outcomes, and how will we identify them?

ACTION

What action will be done if it is predicted that the development or deployment of the algorithm may result in undesirable outcomes?

HOW WILL POTENTIAL BIAS BE DETECTED?



How and when will the algorithm be tested?

Who will be the targets for testing?

What will be the threshold for measuring and correcting for bias in the algorithm, especially as it relates to protected groups?

DESIGN, TESTING, AND IMPLEMENTATION

Questions to Consider

INFRINGED FUNDAMENTAL RIGHTS

Are the design testing and implementation going to impact fundamental rights such as, for example, privacy and data protection, freedom of expression, effective remedy and due process, rights to protection against discrimination, the right to explanation, access to information, freedom of religion, freedom of association, and other fundamental rights as defined by the International Bill of Rights and national human rights law?

SPECIFIC LEGISLATION

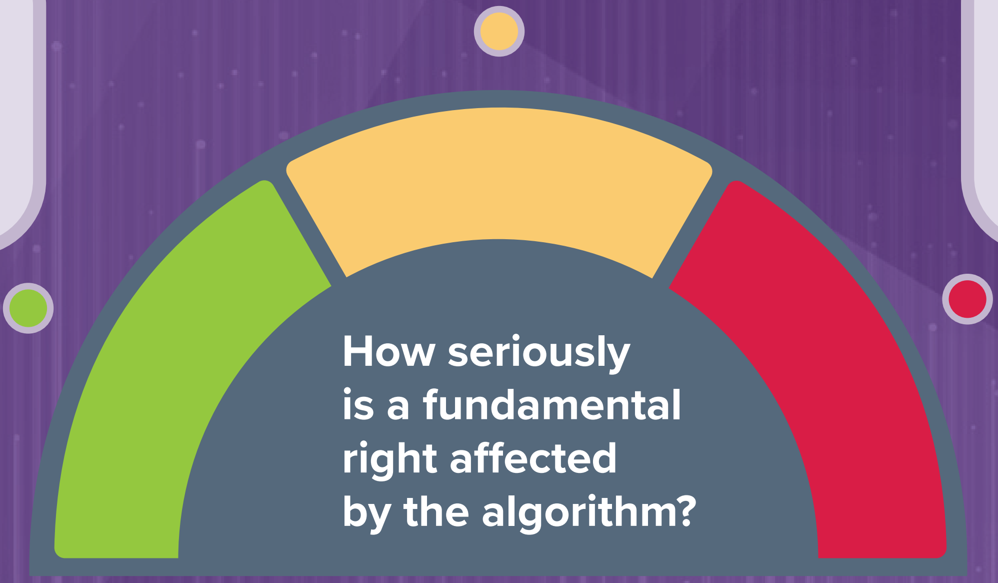
Is there a specific legislation that limits the design, testing, and implementation of the algorithm?

SERIOUSNESS OF INTERFERENCE

Less serious interference, thus no special due diligence required.

Medium-serious interference, thus due diligence required.

Serious interference, thus compelling reasons required as justification.



How seriously is a fundamental right affected by the algorithm?

A useful risk based assessment framework is provided by the EU AI Act.

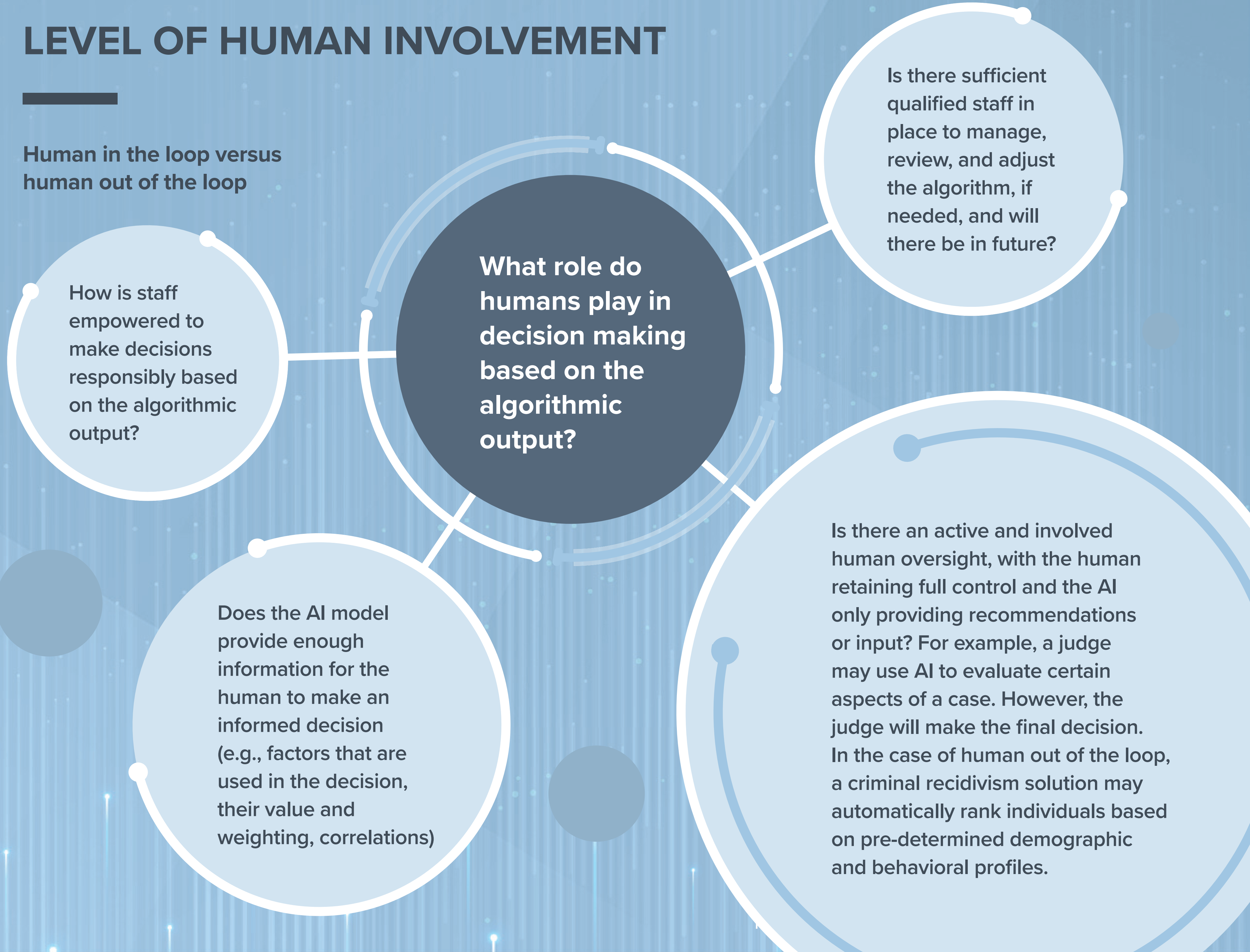
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex-%3A52021PC0206

MONITORING AND EVALUATION

Questions to Consider

LEVEL OF HUMAN INVOLVEMENT

Human in the loop versus human out of the loop



What role do humans play in decision making based on the algorithmic output?

How is staff empowered to make decisions responsibly based on the algorithmic output?

Is there sufficient qualified staff in place to manage, review, and adjust the algorithm, if needed, and will there be in future?

Does the AI model provide enough information for the human to make an informed decision (e.g., factors that are used in the decision, their value and weighting, correlations)

Is there an active and involved human oversight, with the human retaining full control and the AI only providing recommendations or input? For example, a judge may use AI to evaluate certain aspects of a case. However, the judge will make the final decision. In the case of human out of the loop, a criminal recidivism solution may automatically rank individuals based on pre-determined demographic and behavioral profiles.

INTERNAL PROCESS SAFEGUARDS

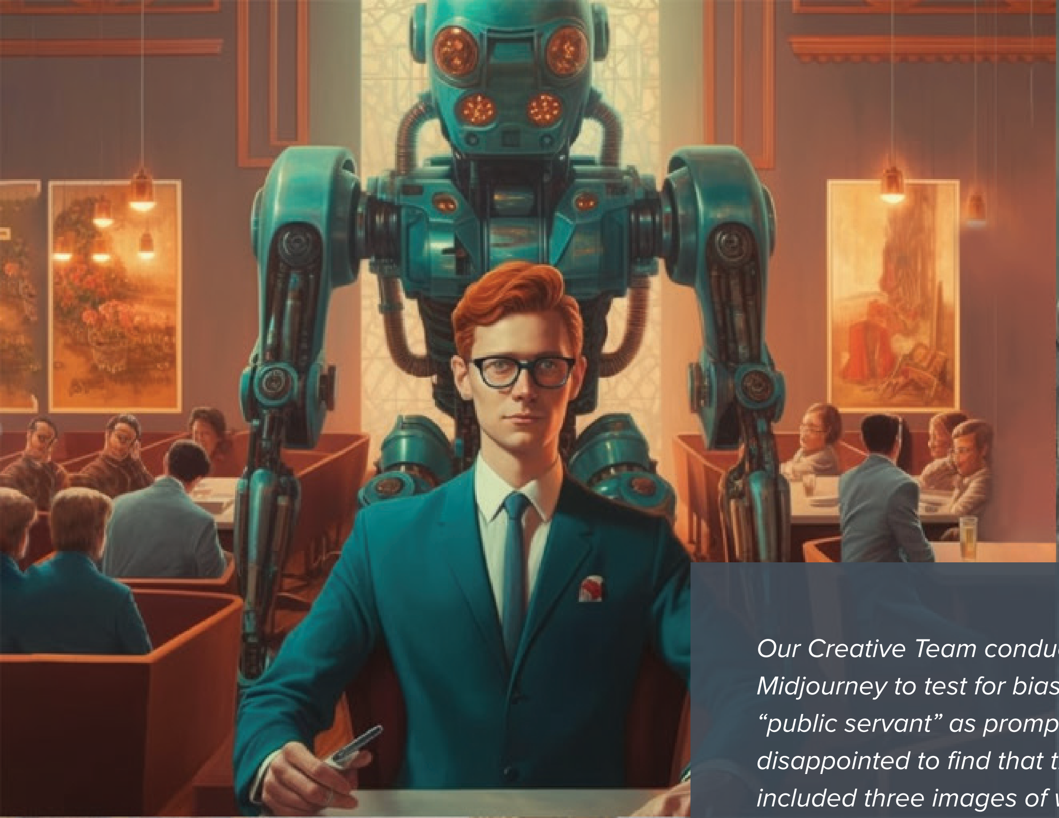
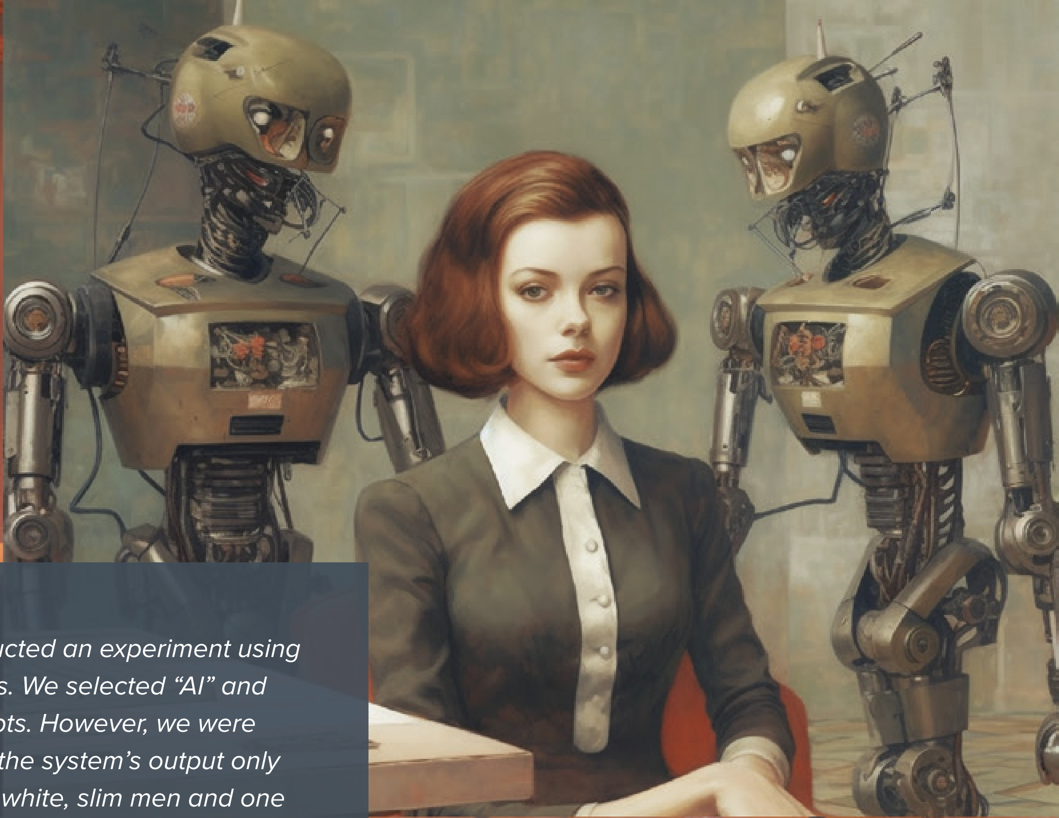
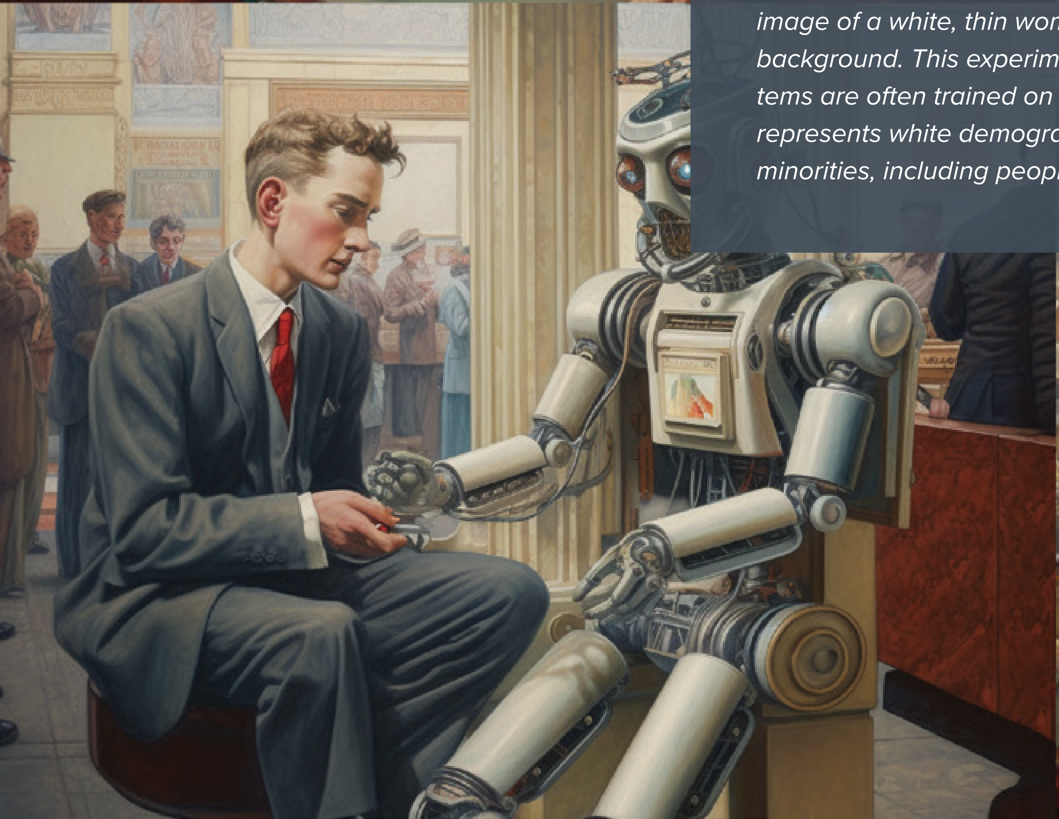
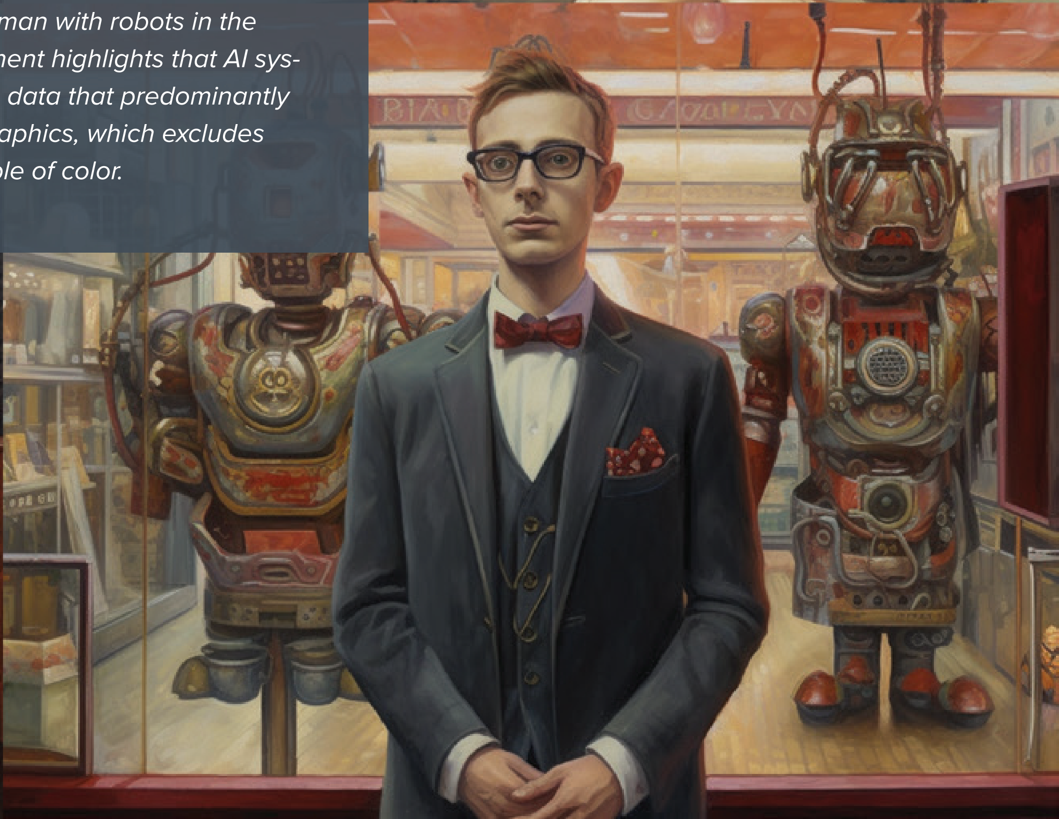
Does the organization have the right staff in place to do so?

How often and at which moments in time is the use of the algorithm evaluated?

Regarding self-learning algorithms: have processes and systems been set up to monitor models (e.g., with respect to data drift, concept drift, and accuracy)?

EXTERNAL PROCESS SAFEGUARDS

Is there a mechanism for external auditing and supervision in place?

Our Creative Team conducted an experiment using Midjourney to test for bias. We selected "AI" and "public servant" as prompts. However, we were disappointed to find that the system's output only included three images of white, slim men and one image of a white, thin woman with robots in the background. This experiment highlights that AI systems are often trained on data that predominantly represents white demographics, which excludes minorities, including people of color.