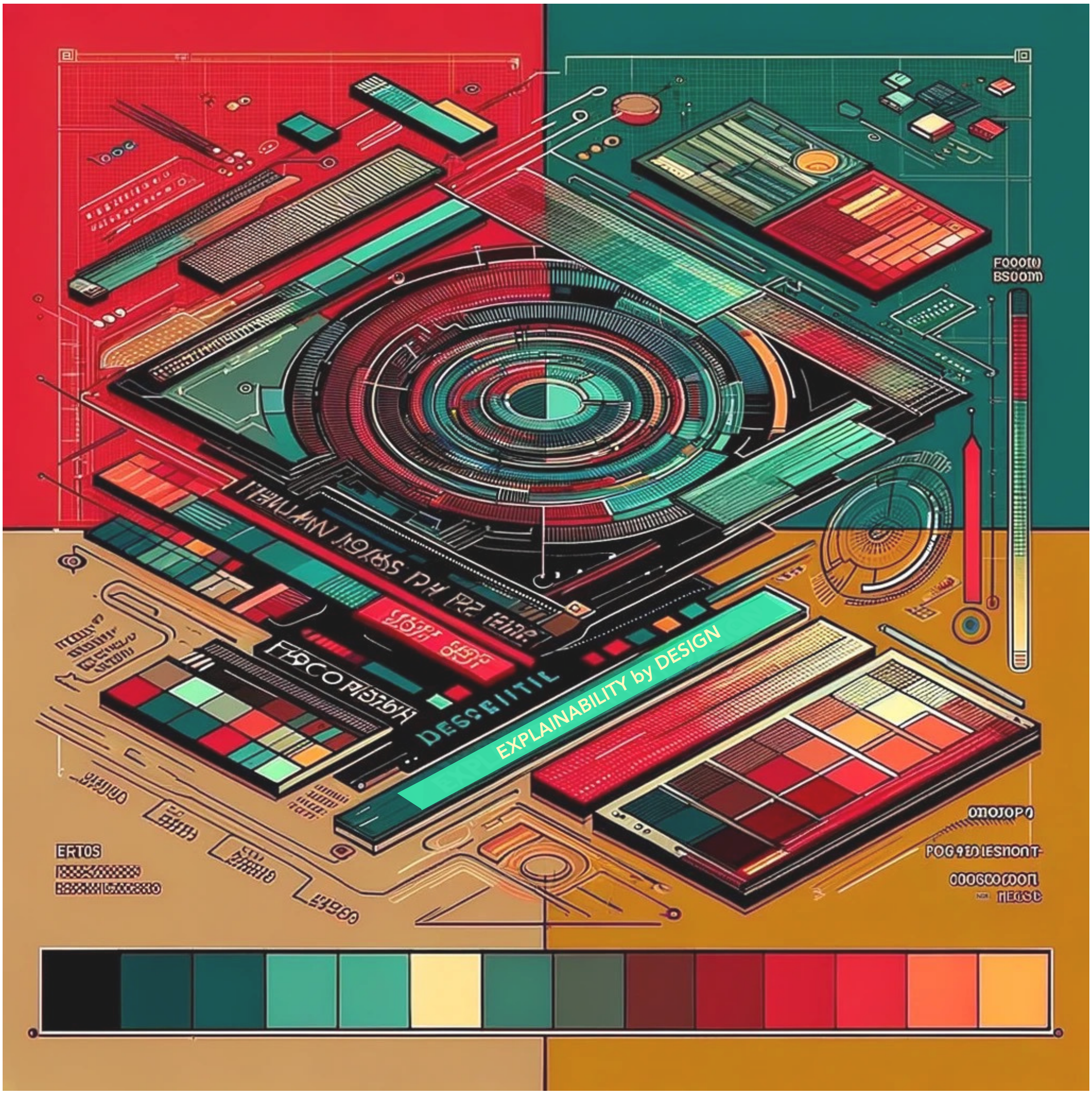
EXPLAINABILITY BY DESIGN

The landscape is changing but the horizon is clear.*



Hello : Dall-E (The AI model that created this image.)

*This is something I have said in podcasts and interviews and meetings for many years with regard to Privacy, but which also applies to AI -from a regulatory standpoint- and, to my mind, with regard to the very real inevitability of, and need for, Explainability by Design.

Shoshana Rosenberg

EXPLAINABILITY BY DESIGN

It is our collective responsibility to ensure that, as AI systems grow more complex and integral to our existence, they do so only where they are understandable, accountable, and aligned with our fundamental values. This approach is essential to uphold **Privacy**, foster **Diversity**, and ensure **Equity** and **Inclusion** in an AI-driven future. ***Explainability by Design*** is the means by which we can harmonize AI development with the requisite ethical imperatives and applicable practical context.

As Artificial Intelligence (AI) and AI regulations will continue to issue and evolve, changing the both the landscape ahead and the path forward, the horizon is inevitably and unmistakably shaped by the concept of Explainability by Design. The goal of this endeavor is to present what I hope is a sufficiently refined straw man set of principles to add to both the conversation and to the call to arms for explainable AI- reinforcing the need not only to break open the black boxes, but to be building auditing and explainability into AI systems from the start.

With the Privacy by Design principles and frameworks to lead the way, and an abundance of global thought leadership in AI ethics and transparency, the *Core Principles* and *Special Aspects Principles* set out here stand on the shoulders of giants, and the wisdom of so many, and are meant to take a stab at harmonizing something for discussion and debate at a more grassroots level. These principles, collectively, form the underpinnings of a proposed approach to Explainability by Design. The *Core Principles* set the stage for ethical and responsible AI, while the additional sets - focusing on development, evolution, and audit accountability - delve into the specifics, taking aim at finding ways to ensure that AI systems are not only ethical by design but remain so through their lifecycle.

From DARPA's pioneering efforts in creating interpretable AI models to the European Commission's Ethics Guidelines for Trustworthy AI, each set of foundations already set out contributes a vital piece to the tapestry of what is required for explainability. The *Core Principles* align with the OECD's emphasis on AI transparency and accountability, the AI Now Institute's advocacy for robust audits and public engagement, and the IEEE's Ethically Aligned Design, which prioritizes human well-being and inclusivity as well as technical integrity.

Explainability by Design is ket to the development of AI systems that uphold privacy, diversity, equity, and inclusion, but it is also key to ensuring that the people and organizations that use these tools can themselves be transparent and accountable. This pass at setting out principles for discussion and refinement is very much one with collective origins and is meant to reflect and facilitate the shared global endeavor at hand.

Explainability by Design is an acknowledgment both that the path to ethical AI is continuous and ever-evolving, and that all regulatory roads lead right back to the start of building (or building with) these tools.

As global citizens, as privacy and governance practitioners, regulators, and innovators, we have the chance to help shift the terrain ahead not just by calling for, but by providing the foundations that will jumpstart, the AI future that is ethical, equitable, and transparent. This is our collective challenge and opportunity. Whatever your specialization or field, I feel strongly that we are all needed to help define and shape the future of AI.

Shoshana Rosenberg

EXPLAINABILITY BY DESIGN
CORE PRINCIPLES

TRANSPARENCY AND COMPREHENSIBILITY
AI systems must offer clear, understandable explanations of their operations, data processing, training, decision making and data provenance, tailored to addressing and anticipating the needs of diverse users.

DOCUMENTATION AND TRACEABILITY
There must be comprehensive documentation of AI systems, including their development process, data sources, decision-making criteria, and changes over time.

AUDITABILITY AND INTERPRETABILITY
AI systems and their decision-making processes, feature importance, weighting, and fairness metrics should be built to be auditable. Mechanisms should be in place to interpret complex data and decisions and made accessible and intelligible to both human auditors and advanced tools for thorough analysis.

ACCOUNTABILITY AND RESPONSIBILITY
AI developers and operators must take responsibility for their AI systems' decisions, and for ensuring ethical and legal compliance. Effective mechanisms must be in place for addressing and rectifying harms.

STATEMENTS

Shoshana Rosenberg

ADAPTABILITY AND EVOLUTION
AI systems must be designed to adapt and evolve in response to new requirements, limitations, and regulatory changes.

HUMAN OVERSIGHT AND INTERVENTION
There must always be an option for meaningful human oversight, ensuring AI decisions can be reviewed and intervened when necessary.

ETHICAL AND SOCIAL IMPACT ASSESSMENT
Regular assessments must be conducted to evaluate the social and ethical impacts of AI systems, ensuring their positive contribution to society.

ROBUSTNESS AND RELIABILITY
AI systems must be reliable and robust, performing accurately under diverse conditions and safeguarding against manipulation or errors.

FAIRNESS AND NON-DISCRIMINATION
AI must proactively address and mitigate biases, promoting fairness and preventing discrimination.

DATA PRIVACY AND SECURITY
Aligning with GDPR principles, AI systems should also be Privacy by Design and must ensure the highest standards of data privacy and security

EXPLAINABILITY BY DESIGN

OPERATIONAL PRINCIPLES

Operational Transparency

AI must be built to allow for clarity and openness in decision making processes.

(I propose universal tagging conventions, but more on that another day…)

Obstacles and Considerations

Complex algorithms can obscure understanding; proprietary models may limit transparency.

Proposed Solutions

Developing transparent AI models; promoting open-source explainable AI frameworks where feasible.

Precision in AI Explanations

AI systems must deliver precise and dependable explanations. clarity and accuracy, and adaptive explanation frameworks for varied user contexts.

Obstacles and Considerations

Complexity of models can make precise explanations challenging; varied audience understanding.

Proposed Solutions

Layered explanation approach to inform all ranges of users; employ interpretability tools (such as LIME, SHAP); ensure that data used for training is high quality and audited for biases; Train developers on explainability.

Accessible Insights

AI explanations must be understandable across all user groups.

Obstacles and Considerations

Diverse technical foundations and backgrounds of users; potential for information overload.

Proposed Solutions

Layered explanation models; user feedback to refine explanation clarity.

Contextual Clarity

AI must provide or allow for relevant and specific explanations in line with the use case.

Obstacles and Considerations

Dynamic application environments; varying data sets.

Proposed Solutions

Context-aware AI models; regular updates to explanation mechanisms.

Real-Time Clarity

AI needs to provide or facilitate the provision of clear and timely explanations in dynamic settings and concurrent with its decision-making.

Obstacles and Considerations

Balancing speed with explanation clarity; managing real-time data flux.

Proposed Solutions

Optimizing AI models for speed; employing edge computing for faster processing.

Shoshana Rosenberg

EXPLAINABILITY BY DESIGN
DEVELOPMENT AND EVOLUTION PRINCIPLES

Agile Evolution

Continuous improvement and adaptability must be established as integral toAI systems.

Inclusive AI Development

AI development, design, and testing processes must be inclusive, considering diverse perspectives and impacts.

Obstacles and Considerations
Incorporating a wide range of perspectives; overcoming inherent biases.

Proposed Solutions
Diverse development teams; inclusive user testing practices.

Scalability and Performance Optimization

Explanability mechanisms must be built to ensure that as the model evolves, they scale effectively while maintaining or improving performance.

Interdisciplinary Collaboration

Collaboration should be encouraged across various fields and disciplines in the development of AI.

Obstacles and Considerations
Integrating diverse methodologies and goals; effective communication across disciplines.

Proposed Solutions
Interdisciplinary teams;
shared platforms for collaboration
and knowledge exchange.

User Feedback Integration

User input and feedback must be a component factored in to all levels of ongoing development of AI systems.

Shoshana
Rosenberg

EXPLAINABILITY BY DESIGN

AUDIT AND ACCOUNTABILITY PRINCIPLES

Transparent Reporting

AI systems should be auditable and audited for compliance, ethical considerations, and operational integrity. Audit results (at least in summary) should be disclosed to stakeholders and the public, where appropriate, to maintain trust and accountability.

Accountability for AI Decisions

It should be clearly defined who is accountable for the decisions made by AI systems, especially where AI decision-making is autonomous or semi-autonomous.

Documentation and Traceability

Comprehensive documentation of AI systems, including their development process, data sources, decision-making criteria, and changes over time. (Also a Core Principle.)

Explainability Auditability

Like AI decision-making, AI explainability should auditable, enabling comprehensive review.

Obstacles and Considerations

Complex AI architectures;
ensuring comprehensive audit trails.

Proposed Solutions

Standardized audit protocols;
transparent logging of decision-making processes.

Continuous Oversight

AI systems should include regular monitoring and updating to ensure they remain aligned with ethical, legal, and operational standards.

Shoshana Rosenberg

EXPLAINABILITY BY DESIGN

ETHICAL AND COMPLIANCE PRINCIPLES

Ethical AI Framework

Ethical considerations must be built into every stage of AI development.

Obstacles and Considerations
Defining universal ethical standards; balancing ethical considerations with technical objectives.

Proposed Solutions
Ethical guidelines for AI development; interdisciplinary teams for ethical reviews.

Regulatory Adherence

AI systems should be Privacy by Design and built to be in compliance with legal standards.

Obstacles and Considerations
Evolving legal frameworks; international regulatory variances.

Proposed Solutions
Regular legal reviews; flexible system architectures to adapt to legal changes

EXPLAINABILITY BY DESIGN

ENVIRONMENTAL AND SOCIAL RESPONSIBILITY PRINCIPLES

Eco-Conscious Computing

Minimizing the environmental impact must be a consideration in AI system development.

Obstacles and Considerations
Balancing performance with environmental sustainability; measuring carbon footprint.

Proposed Solutions
Energy-efficient computing technologies; environmental impact assessments.

Standards Adherence

Compliance with established standards and best practices in AI development and deployment.

Shoshana Rosenberg

A.I., Privacy and DEI have a high level of interdependence and interconnectedness and are continuously evolving.

All three are tied directly to ethics, fundamental human rights, the future of work, and decision making and bias, which means:

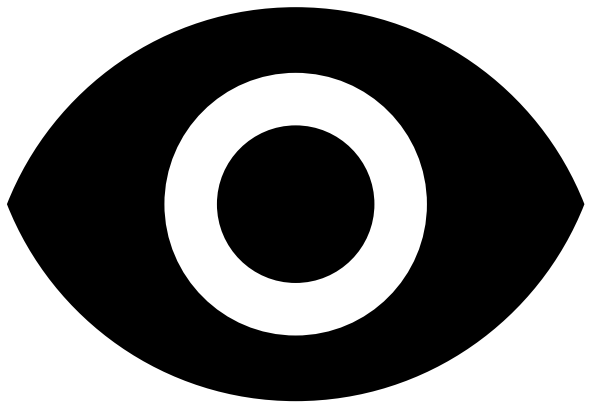
YOU AREN'T ON THE SIDELINES OF THESE THINGS.

YOU ARE CRUCIAL TO THEM BEING WHAT THEY SHOULD.

Do you want to know more?

STATEMENTS

Shoshana Rosenberg



A.
DE.
TEMENTS
STATE
PRIVACY

Shoshana Rosenberg