Dialing with Explainability

Explainability doesn't automatically make ethical AI.
Explainability dials ethics.

What is explainability?

- Code with in-line comments
- Confidence scoring
- Documentation
- And more

Explainability is anything that makes an AI's decisions understandable by people.

Explainable AI can be understood by the people who use it and are impacted by it.

Not understandable

It's technical for a non-technical audience or provides only general information, non-specific to audience.

Built for its audience

It anticipates audiences' questions, needs, and knowledge and adapts explainability to them.

Fosters misguided trust

Through ill advised approximations or otherwise, AI seems more trustable or accurate than it is.

Sets expectations

It empowers audiences to disregard the AI by providing an assessment method for the AI, such as confidence scoring.

Leaves the wrong impression

It omits or simplifies information and causes audiences to believe something incorrect about the AI.

Cultivates the mental model

It enables an accurate perception of the AI and balances detail against the audience's knowledge and biases.