Building "Responsible AI":
Best Practices Across the Product Development Lifecycle

Susannah Shattuck
Head of Product, Credo AI

credo | ai

Nice to meet you, I'm Susannah.

🧑‍💻 Head of Product, Credo AI

🏢 Formerly IBM, Google X, Arthur AI

🧠 Algorithmic bias detection and mitigation, setting up governance structures within AI development organizations, AI regulation & public policy

😍 Speculative design, science fiction, cooking

@shshattuck

The six key tenets of Responsible AI.

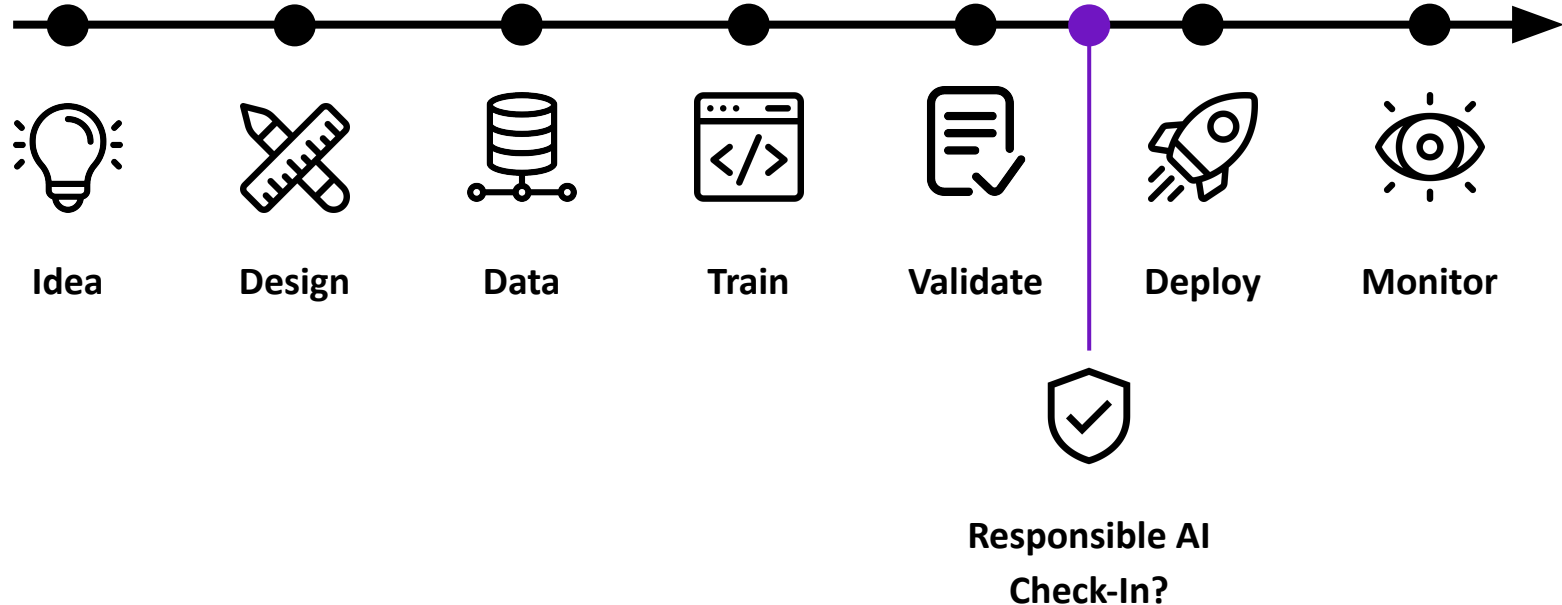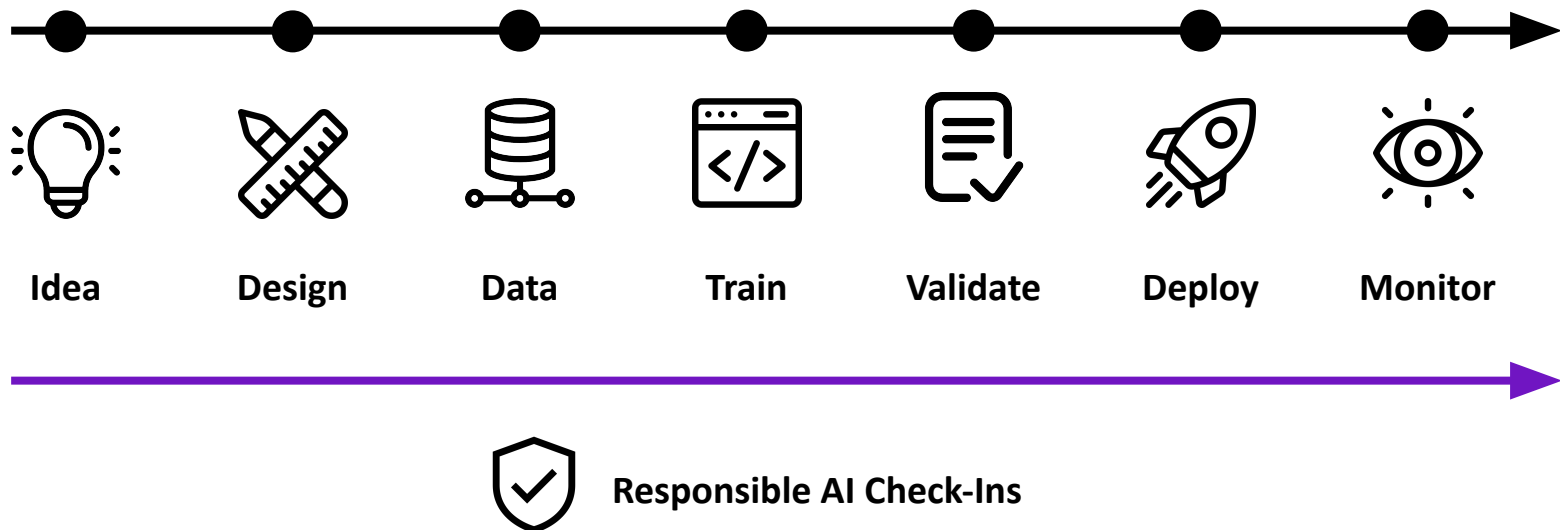|		
FAIRNESS	**TRANSPARENCY**	**SAFETY & SECURITY**
PRIVACY	**SOCIAL & ENVIRONMENTAL WELL-BEING**	**ACCOUNTABILITY**

Responsible AI considerations need to be integrated into the ML development lifecycle.

How does Responsible AI assessment fit into the ML development lifecycle?

Idea   Design   Data   Train   Validate   Deploy   Monitor

Responsible AI

Check-In?

How does Responsible AI assessment fit into the ML development lifecycle?

Idea	Design	Data	Train	Validate	Deploy	Monitor

Responsible AI Check-Ins

How does Responsible AI assessment fit into the ML development lifecycle?

Idea	Design	Data	Train	Validate	Deploy	Monitor
Use Case Review	Align on RAI Requirements	Dataset Assessments	Model Assessments	Risk & Compliance Review	Model Cards, Disclosures	Ongoing Audits

Tl;DR—you need to evaluate the "responsibility" of your AI system at every step of the development lifecycle.

🎨 **During Design:** identify potential risks of the use case and how to measure them

🛠️ **During Development:** prioritize Responsible AI metrics during training and testing

🚀 **During Deployment:** monitor Responsible AI metrics, conduct regular audits

💁 **This is not something that you can or should do alone!** Getting input from different perspectives is key—Responsible AI is a multi-disciplinary problem.
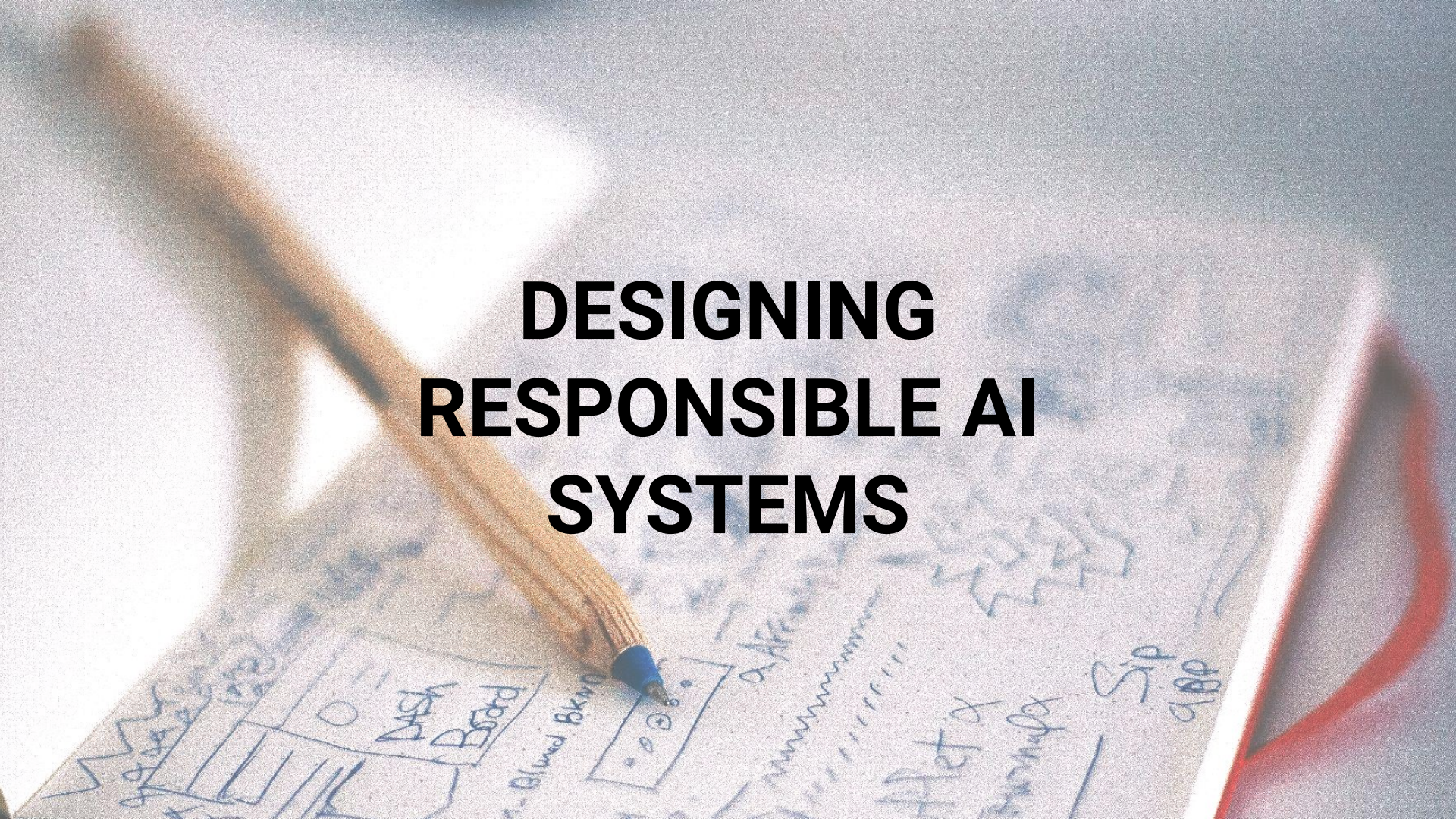
Evaluating your use case: a multi-stakeholder project.

👤 **Who is going to be impacted?** Think about both direct and indirect users; identify all of the groups that will be affected by use of your AI system.

☹️ **What are the potential negative impacts on these people/groups?** Talk to people. Do real user research. Invite impacted groups to participate in the design process.

⚖️ **What is the regulatory context?** Are there any rules, regulations, or standards that need to be followed based on your use case?

📏 **How might we measure and mitigate negative impacts?** Develop a Responsible AI Assessment Plan that will address negative impacts and regulatory requirements.

Tools that help with Responsible AI Alignment:

- AEQUITAS Framework
- Industry standards and benchmarks (NIST, IEEE, etc.)
- Credo AI
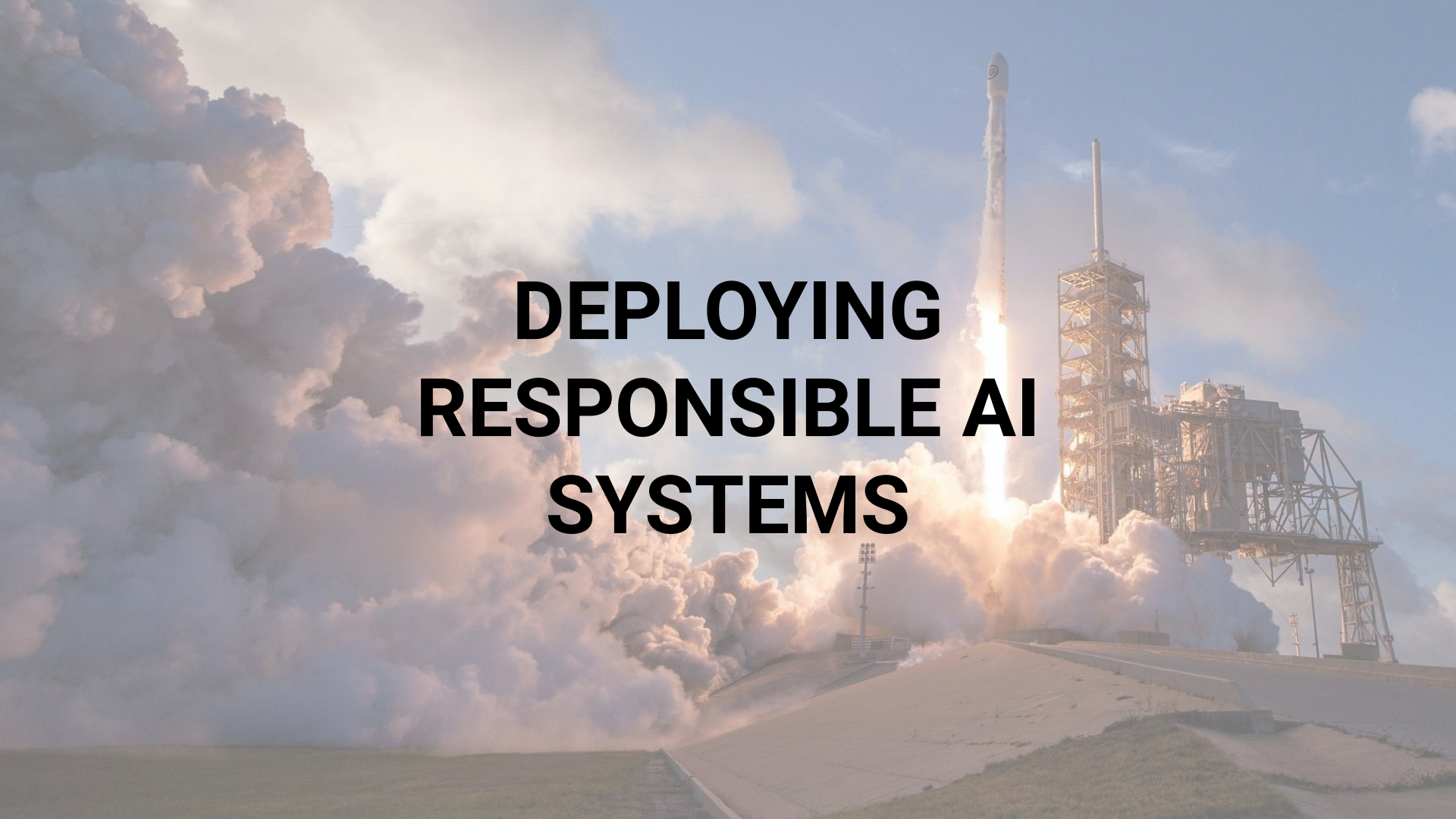
DEVELOPING RESPONSIBLE AI SYSTEMS

Measuring Responsible AI during, not after, development.

🎯 **Include Responsible AI metrics in your objective function.** Optimize for the most performant model that meets your RAI requirements.

📊 **Don't just evaluate your models; evaluate your data.** Fairness and privacy assessments should happen at the dataset level *before* the model level.

🛑 **Rule out model methodologies that don't meet requirements from the start.** Is explainability a regulatory requirement? Don't waste time building a black box model.

📄 **Document your development decisions.** Transparency and accountability are made possible by good documentation; create consistent artifacts during development.

DEPLOYING RESPONSIBLE AI SYSTEMS

Continue monitoring and managing Responsible AI in production.

🔍 **Include Responsible AI metrics in your monitoring plan.** Don't just monitor performance and drift; make sure you're tracking fairness metrics, too.

📅 **Conduct regular stress tests and audits.** Regulations are increasingly requiring regular audits or reports on AI systems' behavior over time.

📢 **Build Responsible AI feedback mechanisms.** Get feedback from your users and the communities impacted by your AI system—and act on that feedback regularly.

🕵️‍♀️ **Have a plan in place if something goes wrong.** Who is responsible for fixing a problem, when it arises? What is your mitigation plan for Responsible AI issues?
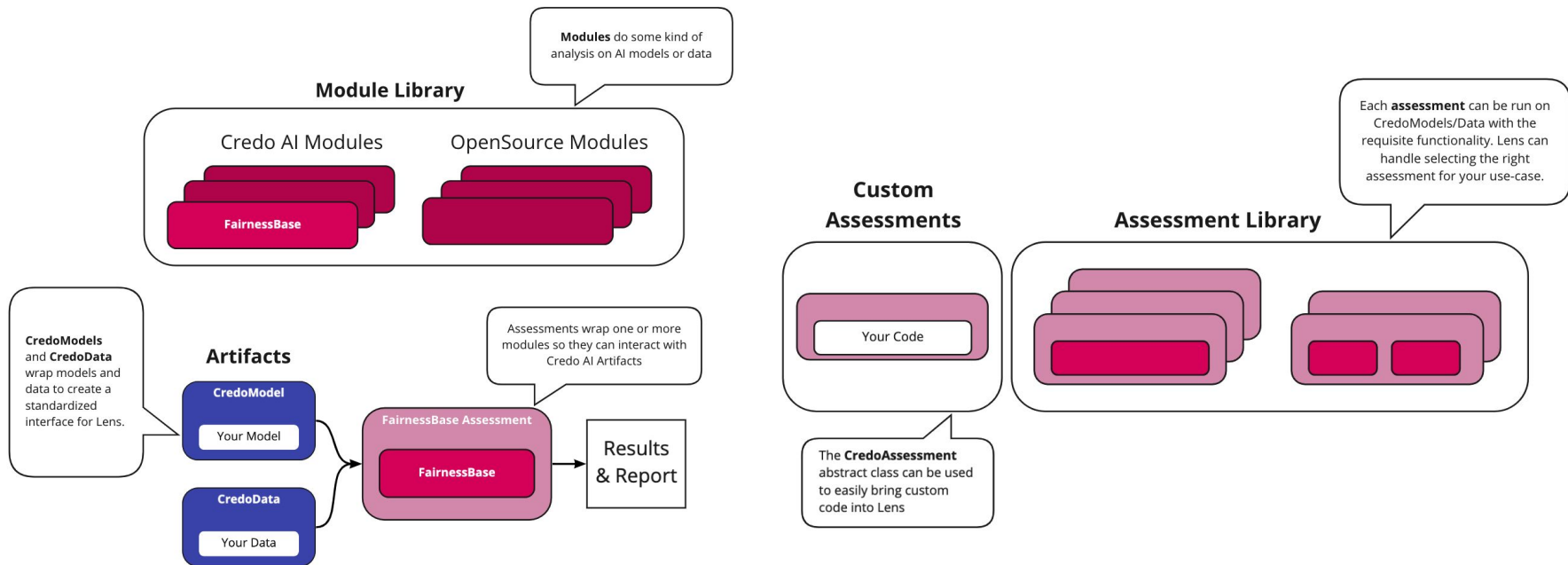
CREDO AI LENS

What you can't observe, you can't control.

Bringing Responsible AI Assessment tools together.

Current Credo AI Lens Assessment Capabilities:

⚖️ **Fairness assessments.** Easily assess parity metrics like disparate impact, equal opportunity difference, etc. for binary classification models.

📈 **Dataset assessments.** Detect proxy variables for protected attributes and get demographic parity analysis of your datasets.

📝 **Custom NLP assessments: toxicity, profanity, verbosity.** For large language models, run a variety of NLP-specific assessments to identify negative model behavior.

👨‍👩‍👧 **Disaggregated performance assessment.** Easily compare disaggregated performance of your model across groups of interest.

https://credoai-lens.readthedocs.io

Thank you!

@shshattuck