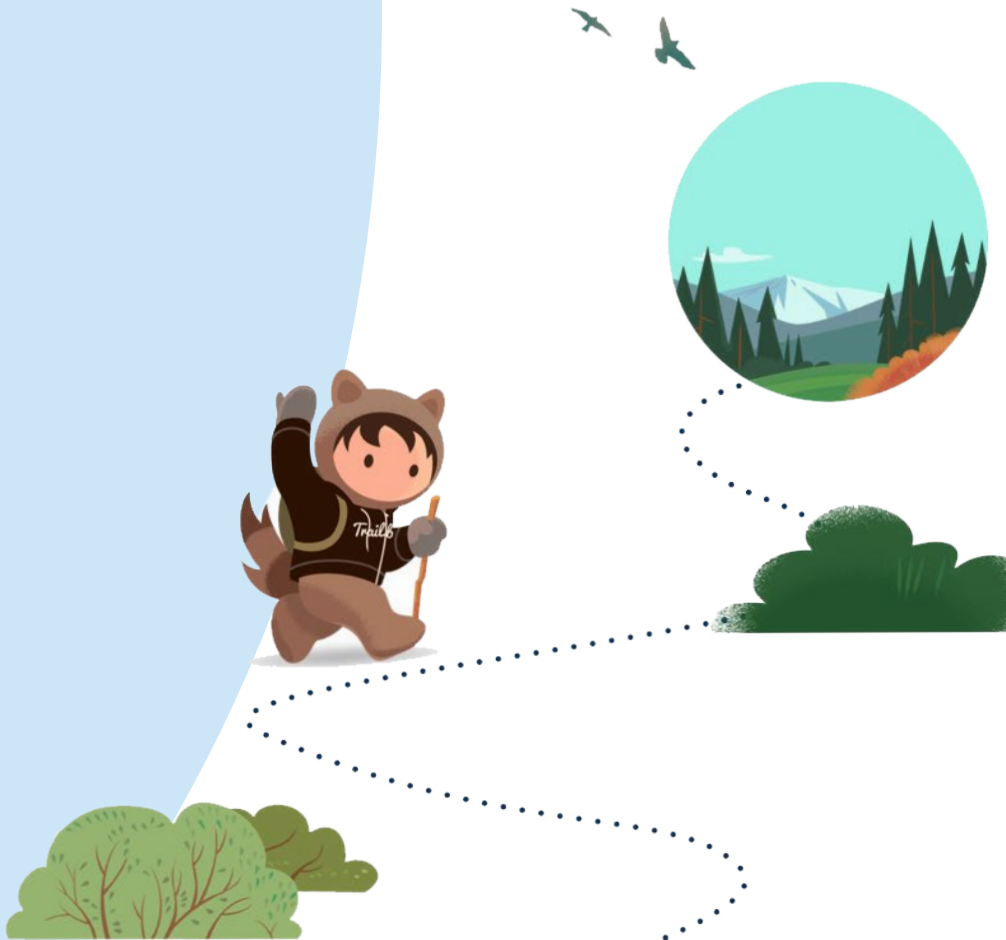
AI Ethics Maturity Model

By Kathy Baxter, Principal Architect, Salesforce Ethical AI Practice

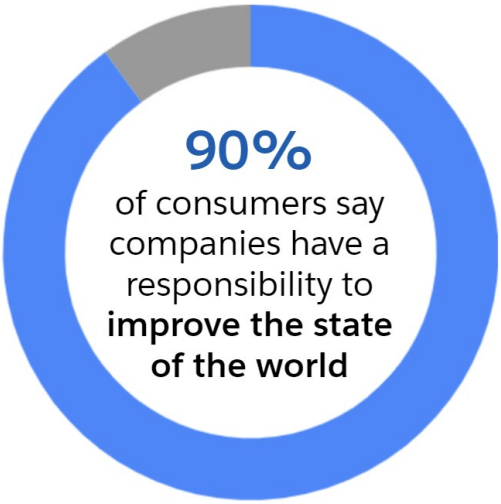
kbaxter@salesforce | @baxterkb



Introduction

Consumers don't trust AI systems but they expect companies to use them responsibly

Research shows that 90% of consumers believe that companies have a responsibility to improve the state of the world. There is guidance for how companies can responsibly create and use technology but many consumers are still concerned about how companies are implementing technology. For example, a global survey in March 2021 found that citizens have low trust in AI systems but expect organizations to uphold the principles of trustworthy AI. If your company is creating and/or implementing AI and want to earn your customers' trust in AI while avoiding both brand and legal risk, you need to implement an ethical AI practice in order to develop and operationalize principles like Transparency, Fairness, Responsibility, Accountability, and Reliability. This maturity model lays out a roadmap for how you might do that based on our own experience, as well as that of other tech companies.



Salesforce Ethical Business & Leadership Survey

For the last few years, Yoav Schlesinger and I have thought a lot about how to grow and mature our AI ethics practice at Salesforce. We've spent time in self-reflection and talking to our peers at other large, U.S. enterprise tech companies that have built their own teams and practices. From this, we've identified a maturity model for building an ethical (or "trusted" or "responsible," choose your own word) AI practice.

Ethical AI Practice Maturity Model

Ad Hoc

Someone raises their hand and starts asking not just "Can we do this?" but "**Should we do this?"**

Informal **advocacy** builds a groundswell of awareness

Ad hoc **reviews** and **risk assessments** take place among "woke" teams

Organized & Repeatable

Executive buy-in established

Ethical **principles** and **guidelines** are agreed upon

Build a **team of diverse experts**

Company-wide **education**

Ethics **reviews** are added onto existing reviews, often at the end of the **dev process**

Managed & Sustainable

Ethical **considerations** are baked into the **beginning of product development** and reviews happen **throughout the lifecycle**

Build or buy bias assessment **and mitigation tooling**

Metrics are identified to track progress and **impact** post-market for regular **audits**

Optimized & Innovative

End-to-end inclusive design practices that combine ethical AI product and engineering dev with privacy, accessibility, and legal partners

Ethical features and resolving **ethical debt** are a formal part of **roadmap** and **resourcing**

Poor ethics metrics **block launch**

Ethical AI Practice Maturity Model: Ad hoc, Organized & Repeatable, Managed & Sustainable, Optimized & Innovative

Ad Hoc

Many of the ethical AI teams created in the last 3-5 years were the result of employee advocacy or interrogation of the AI models or applications their companies or teams were building. Some asked about potential bias in data sets being used to do model training, while others saw the output of biased systems like Microsoft Tay or facial recognition systems and asked how that could happen. They also found other like-minded individuals in the AI community via conferences, social media, and meetups to learn from each other. Today more executives and companies are recognizing that unethical AI can result in legal, brand, and financial risk. As a result, more executives are initiating the creation of ethical AI teams.

Salesforce Ethical Business & Leadership Survey

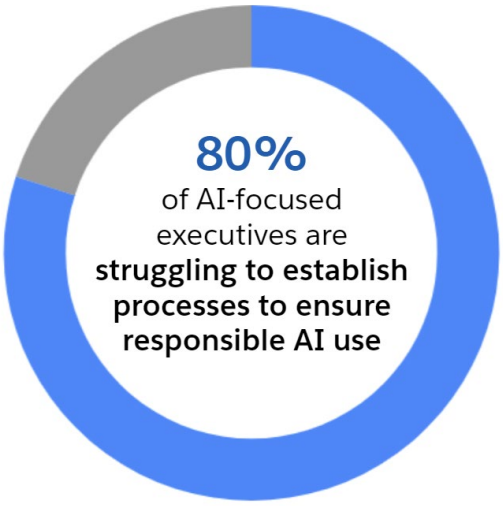
In the ad hoc stage of the maturity model, individuals begin identifying unintended consequences and informally advocating for the need to consider bias, fairness, accountability, and transparency in their companies' AI. And it is this advocacy that creates a groundswell of awareness among other individual contributors and managers to pause and ask not just "can we do this?" but "should we do this?" Creating a discussion group on the company's internal social media channel is a great way to share knowledge, excitement, and identify advocates for the work. Informal tech talks are another way of spreading knowledge and creating dialog within the company. Historically, early advocates for this approach have taken on full-time roles within their companies to build an ethical AI practice.

The process of having this formal role created and filled can take a year or more of building trust among leaders and demonstrating the importance of developing AI responsibly. However, as more executives see the importance of a responsible AI practice, companies without an internal advocate are now looking to hire from outside.

Entire teams and dedicated budgets do not emerge overnight, so ethics reviews by the lone ethics expert are often ad-hoc and limited to individuals or small teams that have bought into the importance of a responsible AI practice. These small successes are critical in building up a portfolio of "wins" and earning more advocates across the company.

Organized and Repeatable

At this stage, executive buy-in has been established and the company is developing a culture where responsible AI practices are rewarded. Part of this culture creation is the development of a set of ethical principles and guidelines. Virtually every company with an ethical AI team - including Salesforce (einstein.ai/ethics) - has published a set of guiding principles. There is significant overlap between the principles published by each company and yet it may require significant time to create alignment amongst key internal stakeholders, to solicit commitment to these principles, and then articulate or publish them publicly. This is an important exercise to create dialog across the company, raise awareness of potential harm from AI systems, situate the conversation in the context of the company's values, and gain true investment. Simply taking a generic set of principles and publishing them on your company website will likely be little more than "ethics washing" and result in minimal change.



FICO: State of Responsible AI: 2021 Report

Salesforce Trusted AI Principles

Responsible

Safeguard human rights and protect the data we are entrusted with.

Accountable

Seek and leverage feedback for continuous improvement. Adhere to regulations.

Transparent

Develop a transparent user experience to guide users through machine-driven recommendations. Communicate how AI was developed and its limitations.

Empowering

Promote economic growth and employment for our customers, their employees, and society as a whole. Give customers tools to use AI responsibly.

Inclusive

Respect the societal values of all those impacted, not just those of the creators.

Salesforce Trusted AI Principles: From Principles to Practice

Responsible	Accountable	Transparent	Empowering	Inclusive
- Work with human rights experts	- Invite customer feedback	- Strive for model explainability	- Build AI apps with clicks not code	- Test models with diverse data sets
-Educate and empower customers and partners	- Engage Ethical Use Advisory Council	- Customers control of their data and models	- Free AI education with Trailhead	- Conduct Consequence Scanning Workshops
- Open development and sharing of research	- External ethics review of high risk AI research papers	- Clear disclosure of terms of use	- Deliver AI research breakthroughs	- Build inclusive teams

It is also during the Organized and Repeatable stage that an actual team is formed. This may be through current employees changing their roles to focus on AI ethics full-time or through hiring external experts. Ideally, this team is composed of many skill sets, backgrounds, and intersectional diversity of race, age, gender identity, sexual orientation, backgrounds, and more. A mix of professional experience in human rights, ethics and philosophy, user research, AI, policy and regulations, as well as data science, product and program management, will also yield better outcomes. Diversity is your superpower because different value systems require different mechanisms for fair decision-making.

*Gartner predicts that through 2022, 85% of **AI projects will deliver erroneous outcomes** due to bias in data, algorithms, or the teams responsible for managing them.*

One other thing to note: The individuals who inhabit these Responsible AI roles should not be evaluated on KPIs like product launches or revenue generation. They should be empowered as neutral evaluators who are not penalized when they identify ethical risks or attempt to prevent the launch of a model/AI application because of those risks. Independence is required for honesty and integrity in your ethical AI practice.

In this stage, it's likely that questions of scale are emerging. You may have a public, company-wide commitment but you probably don't have the resources to ensure that every team building or implementing AI is doing so responsibly. Formal employee education is needed because, like security, ethics is every employee's responsibility, regardless of their job title. Informal tech talks have likely been happening already but now you must think about what the foundational information is that every employee working on AI should know and how to contextualize it across different roles (e.g., engineering, product management, UX) and product teams. You want to ensure that teams are asking the right questions and looping in the AI Ethics team (if you have a centralized team) or the ethics expert on your team (if you do not have a centralized team) for moderate to high-risk use cases but aren't inundating you with low or no risk requests.

*43% of survey respondents believe they have **no responsibilities beyond meeting regulatory compliance to ethically manage AI systems** whose decisions may indirectly affect people's livelihoods*

Through education and outreach, you will likely encounter employees that are passionate about ensuring the responsible creation and use of AI. These employees are likely embedded throughout the company and with deeper training can become your eyes and ears, providing some minimal guidance to others in day-to-day discussions or design decisions.

Another way to scale is by adding ethics reviews to existing AI product reviews. If an ethics review is tacked onto the end of the product development process, right before launch, it leaves teams little time to make significant changes. As a result, this review should happen early. For example, if the training data used are biased and/or not representative of all the users that will be impacted by an AI system, there is little mitigation that can be done to address the potential harm once the model is built and ready to deploy. Although tacking a review on at the end of the development cycle is seemingly the easiest approach and causes the least friction, it is not the most efficient either in terms of harms remediation or deploying development resources. PM and engineering teams hate having to spend resources resolving debt, in this case "ethical debt," accumulated in previous releases.

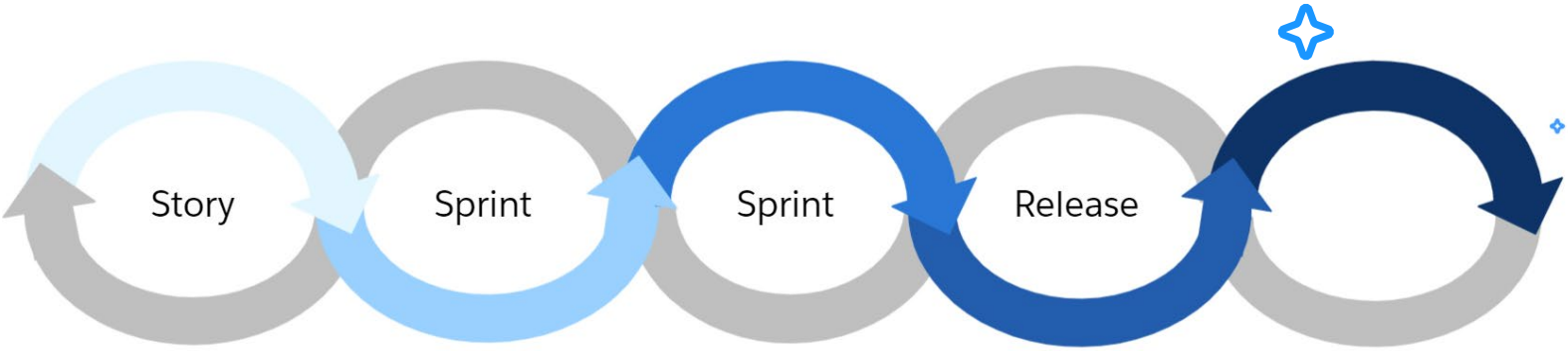
"Ethical debt" is the formal logging of ethics issues identified in a prior release. You may already be familiar with "technical debt;" the cost of additional rework caused by choosing a cheaper/faster/easier solution instead of using an optimal approach that would take longer/cost more. "Ethical debt" is accrued when you launch features that violate your ethical AI principles because, for example, you didn't do a bias assessment or you didn't mitigate the bias that was found. When ethical AI debt is found, it can be far more costly than your standard technical debt because you may have to identify new training data and retrain your model or remove features that you later identified cause harm. Regrettably, it may take a few painful cases of blocking a launch of a potentially serious violation of the company's AI ethics principles in order for the ethics reviews to be added much earlier and throughout the development lifecycle.

Managed and Sustainable

Although it is not perfect, you now have a manageable practice that can scale as the company grows. Depending on the size of your company and success at educating existing employees, you may be able to shift your focus to ensuring new employees know what their role is in ensuring responsible AI. Employees at many companies have a lot of mandatory training to attend so it is worth considering how much training should be mandatory. Every employee working on AI should at least know your ethical AI principles and any customer restrictions on how your AI can be used (for example, at Salesforce, we do not allow our vision AI to be used for facial recognition). Beyond that, deeper training can be limited to a smaller subset of employees. For example, it might be the case that only engineers or data scientists building AI need to know how to quantitatively assess bias and how to mitigate it.

At this point, your company has introduced ethics checkpoints throughout the product lifecycle. Formal processes like consequence scanning workshops, ethics canvas, harms modeling, community juries, and creation of documentation like model cards (like nutrition labels for models) or FactSheets are implemented and required by management. The addition of new processes and documentation will likely grow as your practice matures.

Responsible AI Development Lifecycle



Scope

Should this exist?

What are your assumptions?

Who are you designing for?

Known risks?

Review

Who will be impacted?

Who is excluded?

Consequence Scanning

User Research

Test

Bias Assessment in Dataset & Model

Ethical Red Teaming

Community Feedback

Mitigate

Bias Mitigation

Retrain Model

Compare to Threshold Set for Launch

In-App Support

Launch & Monitor

Publish model cards

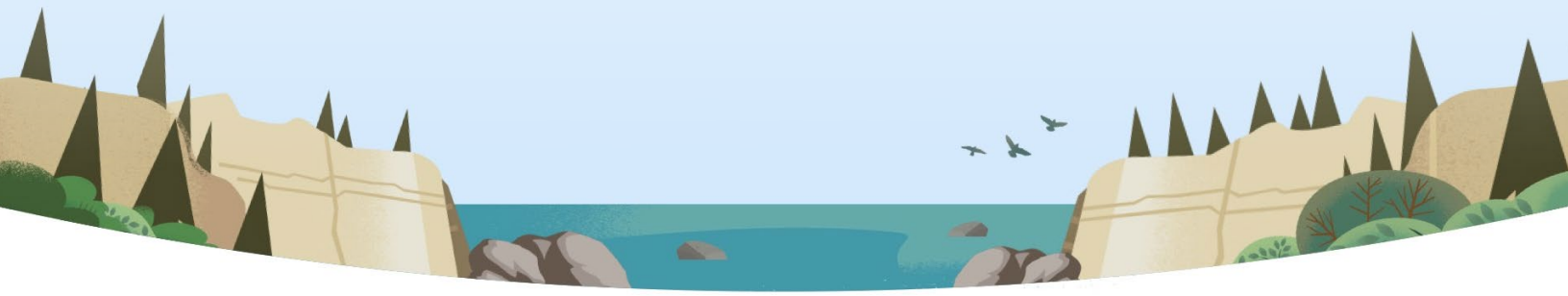
Post-Launch

Assessments

Logging

Community Feedback

Responsible AI Development Lifecycle following the Agile Development Lifecycle stages of Story, Sprint, and Release



The use of FATE (fair, accountable, transparent, explainable) tooling and engineering practices by AI engineering and data scientists is typically introduced at this stage in order to consistently identify potential bias and mitigate it in training data and models, as well as to increase the explainability of the models. The output from these tools and practices will inform your model cards or FactSheets.

The other new practice that is introduced at this stage is the establishment of metrics to track the progress of the ethics work and the impact of bias on customers. You can never claim that a data set or model is 100% bias-free or completely fair. Any model that appears to be 100% bias-free will suffer from overfitting and, in all likelihood, poor performance. Instead, you can only share what type of bias you looked for, how you measured it, what you found, and what you did to mitigate the bias and potential harms you anticipate. In order to know if you are making progress and what impact your AI systems are having on your customers, you need to establish metrics to track, a reporting mechanism to publish those metrics, and an incentive structure that rewards continuous improvement.

Applying metrics to your internal processes and models is just the beginning. You also need to conduct post-market monitoring and auditing to understand the real-world impact of our AI on customers and society. Bias and fairness metrics in the lab are only an approximation of what will be measured in the wild. For example, even if your facial recognition technology (FRT) had the same error rate for all genders and skin tones before deployment, the impacts when the system "gets it wrong" can be quite different for some populations. The experiences faced by innocent black men misidentified by FRT will likely be different from innocent white women that are misidentified. You need some way
to monitor for harm and for individuals to ask for redress and remediation.

If you began your ethical AI practice in one country and you have customers outside of that country, you must expand your work to be inclusive of multiple languages, cultural values, and contexts of use. You can't simply overlay what you were doing in the US, for example, onto your customers in Japan or Mexico and expect that you will identify all of the ethical risks or know how to mitigate them. It bears repeating that different value systems require different mechanisms for fair decision-making. The historical or societal bias that exists in training data will differ by region and language and therefore must be examined within the context of those regions and languages. If you haven't already, you need to start hiring in other countries.

Optimized and Innovative

This is the end state you are striving for. But we intentionally refer to our work as a "practice" because the goal is continuous improvement -- there is no such thing as "perfection" in this work. As new AI applications and methodologies are developed, new ethical risks are identified and new ways of mitigating them may be needed.

In order to create end-to-end ethics-by-design, mature AI ethics practices combine ethical AI product development and engineering with privacy, legal, user research, design, and accessibility partners to create a holistic approach to the development, marketing, sale, and implementation of AI. You may also have moved from a large centralized AI ethics team to a hybrid or hub-and-spoke model. In the hybrid model, a centralized ethics team owns standards and the creation of new processes while individual ethicists are embedded in AI product teams to provide dedicated, context-specific, and timely expertise. There is no one "right" model; it depends on the size of your company, the number of product teams building AI applications, how diverse those offerings are, your company's culture, and more.

At this stage of the maturity model, product roadmaps and resources explicitly require that ethical debt is addressed and new features to help customers use your AI responsibly are highlighted. Since metrics were established in the previous stage, it is now possible to set minimum thresholds for launch in order to block the launch of any new product or feature that does not meet that threshold. Of course, we know that metrics can be manipulated so you don't want to depend on a single metric, or even a few, as the sole go/no-go factor in the decision-making process around launch. However, these metrics should be discussed and deeply understood by those deciding what is ready to launch.

Conclusion

The Ethical AI field is relatively new and we are all learning together as we understand risks and harms associated with certain AI technologies or applications of them to different populations. The proposed maturity model will change as our understanding and practice develops and it is our hope that we can co-create this field together.

Additional Resources

AI Ethics Blog Posts

einstein.ai/ethics

Online Learning

Ethical & Inclusive Products Trailmix

sfdc.co/EthicalandInclusiveProductsTrailmix

Salesforce Model Cards

bit.ly/3oNIBMs

Thank You

Kathy Baxter, Principal Architect, Salesforce Ethical AI Practice

kbaxter@salesforce | @baxterkb