Embracing Responsible AI from Pilot to Production

How to Build AI That Works for Everyone

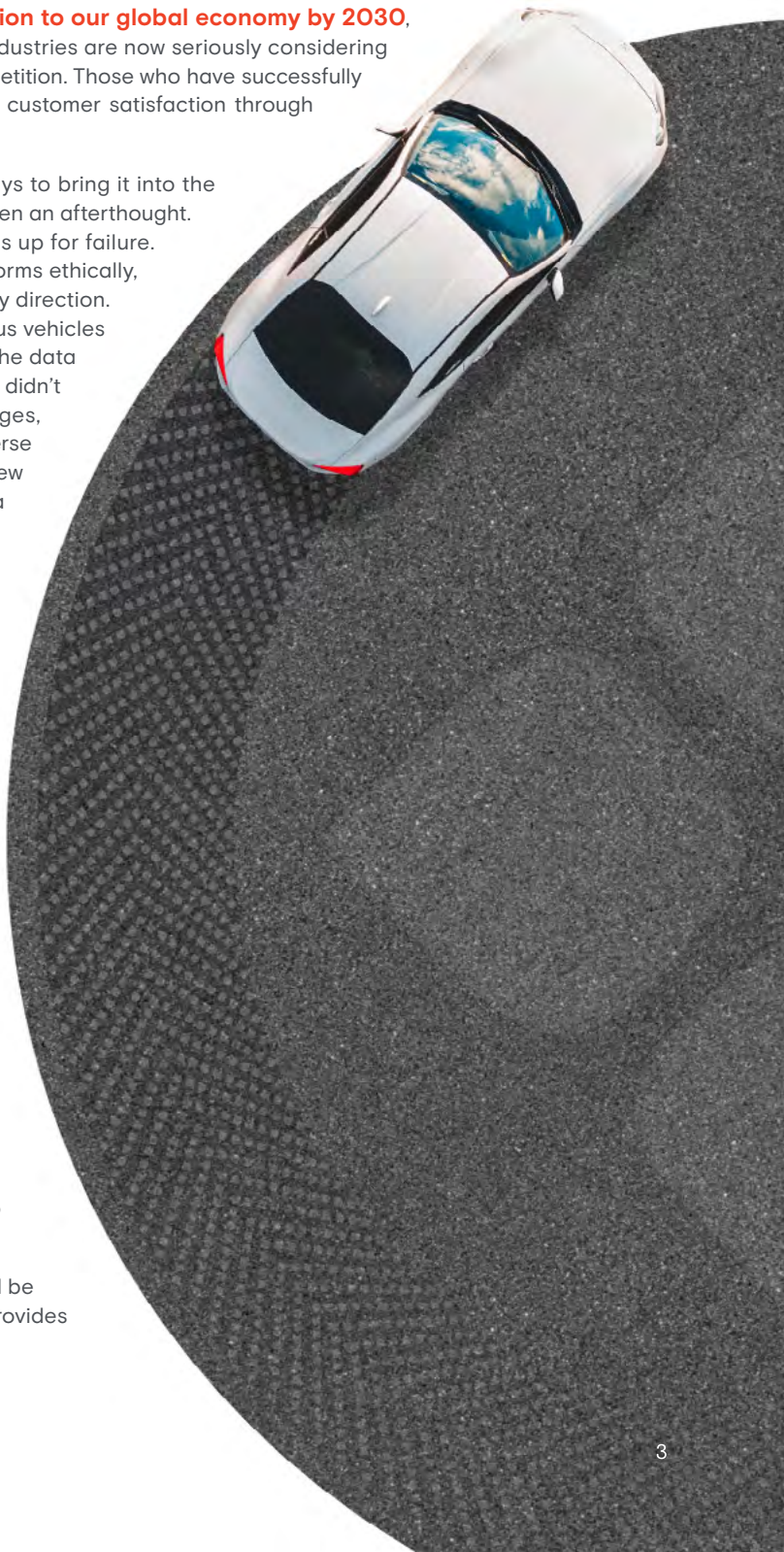
An Appen eBook

appen

Table of Contents

Embracing Responsible AI from Pilot to Production

Introduction



Artificial Intelligence (AI) has the potential to add **$13 trillion to our global economy by 2030**, according to McKinsey. Organizations across all major industries are now seriously considering AI solutions to drive business value and keep up with competition. Those who have successfully deployed AI see high returns on investment and increased customer satisfaction through more personalized, efficient AI tools.

With AI investment growing and teams rushing to find ways to bring it into the core of their business, the intention of responsible AI is often an afterthought. This is one of the biggest ways companies set themselves up for failure. While many people think of responsible AI as AI that performs ethically, it's important to consider responsible practices from every direction. It's easy to believe that a company that trains autonomous vehicles to avoid pedestrians is responsible. But what happens if the data they used to teach their model wasn't diverse enough and didn't recognize people of all different shapes, sizes, colors, ages, etc.? Taking this a step further, what if that data wasn't diverse because the people building the model didn't think to review that the data was inclusive? Or what if they had their data annotated for whatever was best for the bottom line, but perhaps that meant that people were really paid less than a livable wage?

There are numerous important questions around the ethical impact of AI: how do we make AI that works for everyone? How do we mitigate human bias from our machines when humans are building them? Are we being good corporate citizens and responsibly sourcing our data and practicing the right security and compliance with that data? Who has the responsibility to ensure AI solutions are net positive for the world?

Many companies are grappling with how to answer these questions. Still, one thing is clear: **organizations investing in AI committed to building AI that is responsible, ethical, and representative have a better success rate**. There's no tradeoff here: for an AI solution to work, and work well, it must work for everyone. A biased model that works for some users, and not others, is a failed model. Or a model that wasn't sourced responsibly can be a poor reflection of company values and a nightmare with the media. It's helpful to remember that AI reflects the people and the company that build it: when something goes wrong, it shows something may also be wrong internally.

In embracing responsible AI from pilot to production, you'll be able to strategically launch an inclusive AI initiative that provides business value and reflects an ethics-first approach.

What is Responsible AI?

The question may seem simple, but in fact, responsible AI can mean many different things to different individuals and organizations. Some definitions focus more on outcomes—what effect will this AI have on our end-users? Our society? Others are centered more on the policies and processes an organization embeds in their AI pipeline—are these frameworks in themselves responsible and ethical?

Several companies have tackled the challenge of defining responsible AI. For example, Salesforce outlines **five principles for trusted AI**: AI that's responsible, accountable, transparent, empowering, and inclusive.

> **The definition we [at PwC] have is twofold: one is the toolkit, which is a collection of assets both code-based and not code-based that help clients be in control of their AI development across the project lifestyle. The other is the philosophy of doing AI that looks at governance—corporate strategy, procurement, practices, and policies—risk management—how do we identify and mitigate risks—and applied ethics—how to identify values that need to be applied to the product and synchronize them with the values of the organization."**
>
> **— Maria Axente,**
> *Responsible AI and AI for Good Lead at PwC UK*

While definitions are numerous and varied, what remains critical is that everyone in an organization is on the same page on how responsible AI plays a role in their work. Aligning on values creates more collective investment in outcomes. Imbuing those values into every step of an AI project will naturally lead to a more consistent product reflective of the organization's goals.

As the dialogue on this topic continues, we may see increased alignment across organizations on what it means for AI to be responsible. In any case, companies that document their approach to responsible AI are taking a crucial first step in furthering the dialogue and progress in ethical AI.

Challenges of AI Adoption

Only **10% of organizations** achieve meaningful financial benefits with AI, while the remaining 90% are still trying to figure it out. As more organizations attempt to launch AI solutions, patterns are emerging concerning best practices as well as roadblocks to reaching production. In a recent Gartner study, survey respondents selected the factors that proved most challenging to their organization in implementing AI. These were the top three responses:

According to a Gartner study, these are the top three challenges companies face in implementing AI:

- **Skills of staff (56% of survey respondents)**
- **Understanding benefits and uses (42% of survey respondents)**
- **Data scope or quality (34% of survey respondents)**

1. **Skills of staff.** Data shows a gap in machine learning, AI, design, and other relevant skill sets in the technology space.
2. **Understanding the benefits and use cases of AI.** Organizations are struggling to answer questions like "Which business problem can AI actually help solve?" and "How much should I invest to solve this problem?" Getting these answers wrong can be costly.
3. **Data scope or quality.** Organizations have difficulty finding enough high-quality data to train their models, directly impacting the model's accuracy.

There's also emotional fear of what can go wrong that prevents some organizations from diving into AI. AI often has a social impact, and while that means the results can be meaningful and positive, there are also cases where things can go wrong and bias can be introduced. That's why it's crucial to approach AI from an inclusive standpoint.

Fortunately, there are many cases where this has already been done successfully. And many organizations have surmounted challenges around skillset, use cases, and data. Mitigating these challenges is a matter of asking the right questions in your organization, following best practices, and avoiding common pitfalls.

The Five-Step AI Development Cycle

The process of building state-of-the-art, responsible AI can be broken down into five key steps. In each stage, your organization must ask itself important questions that will determine your project's direction and, ultimately, the success of the model you build.

Step 1. Identify Business Problem

In the first stage of the development cycle, narrowly define the business problem you're trying to solve. The key word here is "narrow": if you select too broad of a problem, you risk requiring datasets with far too many labels across an unmanageable amount of data classes. In other words, increasing the scope of the problem will require more resources and complexity, and your organization may not be ready for that. Many organizations fail at implementing successful AI solutions because they didn't select the right problem from the start.

Specific

Measurable

Achievable

Realistic

Time-based

Clarify the goal. A rubric that may help you narrow and clarify the purpose of your AI project is **SMART**: your goal should be **S**pecific, **M**easurable, **A**chievable, **R**ealistic, and **T**ime-based. It should provide a value-add to your business proportional to the time, money, and effort you're willing to put in. Your goal should also align with your organization's values around responsible AI (if your organization has not defined any values specific to AI, start there first). Be sure to identify both the purpose and the limitations of your AI project in fulfilling those values.

Think of your end-user. Explore the business use cases of your AI model. Have your end-user in mind as you brainstorm. Ask yourself:

- Who is my target end-user or customer?
- What traits do these people have?
- Should they all be treated equally? If not, how?
  If so, define equal in what measure.
- In what cases will my AI model benefit my target end-user?
  In what circumstances will it not?

Plan for integration. While forming a picture of your AI solution and end-user, be careful not to overlook the business processes you already have in place, and consider what role your model will play in those. Will your AI solution integrate with existing business processes? If so, how and where? What needs to change about the structure of those processes to accommodate your solution? Do you have the resources to make those changes?

Likely, your AI solution will be integrated with existing processes, but if not, is your organization ready to commit to using it as a standalone component?

Obtain stakeholder buy-in. All stakeholders must align with the organization's values around responsible AI before achieving further alignment on the specific goal or project. Everyone in an organization has the potential to touch AI, be they data scientists and engineers, product managers and financial executives, or legal and marketing. This means it's the responsibility of many functional teams to identify and mitigate ethical risks. Documenting a common understanding of what responsible AI means to the organization and the role team members must play in upholding that definition is a critical foundational piece.

Assuming you've accomplished the initial value alignment step and have now selected a problem that addresses each SMART factor and answered the above questions, you'll want to get buy-in from key stakeholders on the project. Stakeholders will include the teams that are expected to execute the work and invest the resources. Depending on the project's scope, stakeholders may also include members of your organization's executive team and other key members. The more stakeholders you have on board, the better.

Achieving stakeholder buy-in for your business problem will require you to:

1. Identify critical business and consumer use cases for your AI model
2. Share what kind of data you already have, if any
3. Clarify what data you still will need and the estimates of the resources necessary to get there

Through discussions with stakeholders, you should agree on your organization's priority in solving this business problem through your AI solution. This will indicate the level of investment of time, money, and people to complete the project.

Step 2. Gather Data

Congratulations—you've got the green light! If you completed step 1, you're already well ahead of many organizations. Companies will often skip straight to collecting data that is immediately available and then trying to figure out how to use it. Easily-available data tends to be riddled with unreliability and gaps, however. Selecting a problem to solve before data collection aligns your solution more tightly with your organization's needs.

In step 2, you'll want to collect data that is reliable, clean, free of bias, and complete. Using high-quality, representative data is arguably the most critical step in embracing responsible AI and achieving a high-performing model for your end-users, regardless of their age, race, sex, geography, or other differences.

This step shouldn't be taken lightly: **you'll be spending up to 80% of your development cycle on preparing training data**, so plan to invest most of your time and resources here. As you work through this step, answer these questions:

What am I asking my model to do?

Before you dive into gathering data, think first about the decisions you'll be asking your model to make. Could the nature of those decisions create bias?

- **Is your AI making decisions usually made by humans?** There could be bias in that set of decisions because humans are inherently biased. Be aware that the data you select to train your model could reflect those biases, and you may need to expand or change your dataset.

For example, let's say you're training an AI model to determine who should be granted a bank loan. Traditionally, men were more likely to receive loans than women due to greater labor participation, systemic issues, and other factors. An AI model trained on this data may grant more loans to men, even when all other elements are equal—and in fact, this did happen in real life to a married couple with identical finances. Are there ways you could anticipate and correct for something like this happening before deployment?

- **Is the decision your AI making a simple, binary decision – heads or tails?** Think about any context you may have missed. Think about how you can reframe the question to reduce the element of bias in the resulting decision.

For example, if you're asking your AI whether it sees a child in an image, have you thought about developmental disorders that would cause a person to appear younger than they are? How will they be accounted for?

- **Is there a protected class (e.g., race, gender, etc.) or status involved in this decision?** Ensure you're sensitive to any bias or systemic issues inherent to that protected class as you're building your model.

For example, if you are looking to use AI to help speed along recruiting, you'll be dealing with a protected class or status. Even if you optimize your model to work without that specific data point, such as gender, it may be trained with data that is primarily made up of men. This might still bias your recruitment AI model toward selecting more men.

Review these questions ahead of time, and flag any potential issues where things can go wrong. You'll want to mitigate these flags on the front-end to avoid any unfortunate surprises on the back-end.

Next, define the attributes on which you would like your decision to be made. If you're asking your AI, "Is there a human in this image?" you'll need to decide if cartoons count as a yes, for example. Determine what your target data classes should be.

What kind of data do I need?

There are many factors to consider when selecting the kind of data you need. While some elements are specific to the problem you're trying to solve, there are universal guidelines you should follow:

- Your data should be secure. Your data governance processes should also meet the regulatory and compliance requirements of your project. A data partner can offer you security options at the level you need.
- Your data should be clean. You may need to remove any incomplete, incorrect, or irrelevant pieces of data if that has not already been done.
- Your data should be complete. It should cover your business use cases and potential edge cases.

Think about this example: when doing an image search for "nurse," a search engine may provide only images of women, as more women are nurses than men. If you train a model on that data, the model may infer that all nurses are women, which is incorrect. In cases like this, you need to gather additional data or create simulated data covering all use cases.

Where am I going to get my data?

Ideally, you'll source multiple datasets from multiple places to enhance diversity and reduce the chances of bias.

You may choose to use internal data, although most organizations don't have the breadth of high-quality training data required for an at-scale project. Still, if you do go internal, use the most comprehensive datasets available and experiment with different datasets and metrics.

If you leverage external data, sources may include real-world usage data, survey data, public data, and simulated data. Choosing the sources for your data will, of course, depend on availability and what makes sense for your model. Still, care should always be taken to obtain diversified sets representing your end-users and use cases.

For both internal and external data, ensure you have permission to collect and use the data. Knowing where your data comes from and thoroughly understanding your data will enable you to avoid hidden surprises that could impact you in production.

Who will collect and annotate my data?

You'll need a team to collect and label your training data. Depending on your data needs, this can be a momentous task that requires a huge investment of time and people, which is why most organizations work with a trusted data partner. The right data partner will give you access to a diverse group of people who will build out training datasets for your model.

When selecting your data partner, be sure you're choosing a partner who commits to working with you on building responsible AI. At the very least, companies should begin developing an AI ethics strategy that is documented and available for your viewing. In practice, this means one, that they're providing you with a diverse crowd that includes various ages, races, geographies, languages, and gender to mitigate bias. The crowd should reflect the values you want incorporated into your data. Your data partner should also be committed to fair pay and fair treatment of their crowd workers, an often overlooked component of building ethical AI.

Look for a data partner that can offer end-to-end support and expertise, and will work with you on ethical implications throughout the development cycle.

Step 3. Model Build

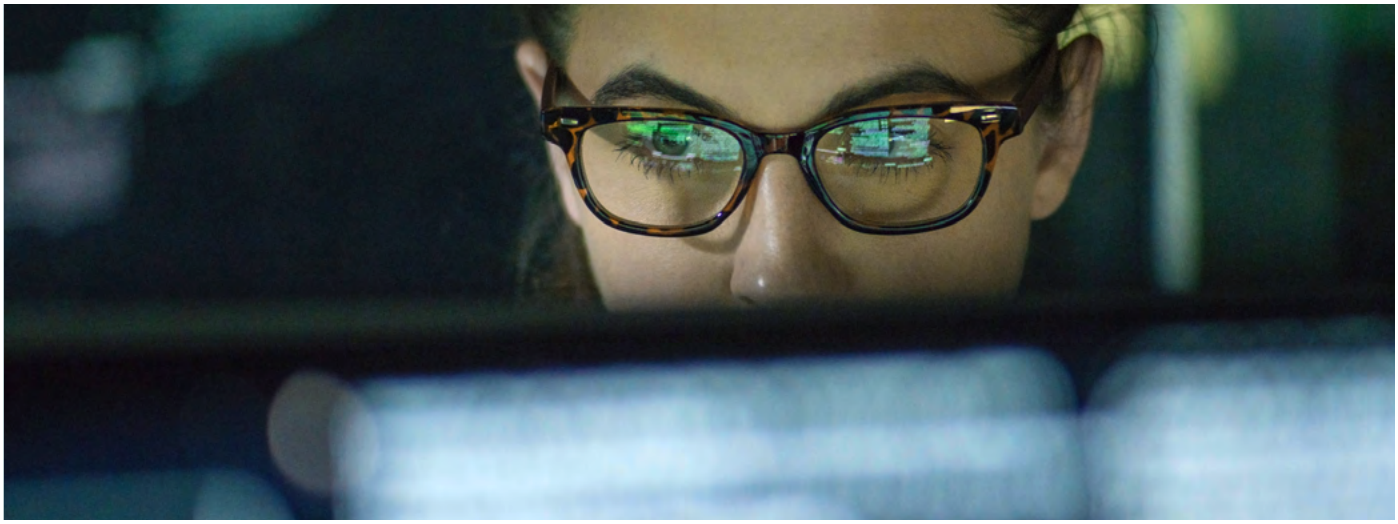
Step 3 is training, testing, validating, and tuning your model. This is your last opportunity to catch bias before your model goes to production. Create space for knowledge sharing and cross-functional feedback loops as you build out your model to mitigate bias further and ensure you're leveraging best practices across your organization.

As you work through these processes, it's critical to document every aspect of your model-build to make changes more efficiently in the future and replicate all or portions of the model in the future.

Training

Allow for differing opinions. Enable your model to give an "I don't know" answer, or provide a low-confidence benchmark so you know when your data, labels, or other parameters need adjustment.

Use a human-in-the-loop approach. Often, companies make the mistake of not including humans as a check-and-balance measure when building their model. Keep people involved at a point in the decision-making process to provide a sanity check (i.e., is the model labeling these pictures correctly? If not, adjust).



Step 4. Deploy and Measure

If your model has successfully passed the testing and validation phases, it's time to scale to production. Before launch, do a final check that the model you created is effectively answering the business problem you identified in step 1. When considering moving to production, if you followed step 1 you should have a plan already on where your model will fit in your existing business processes. Now is the time to complete that integration.

Next, you need to have a tracking mechanism in place to monitor the performance of your model. Remember to include both business and ethics performance measurements. Have a dashboard set-up to look at run-time data continuously and set up alerts for yourself so you have visibility into what's going on at all times. Make sure you are keeping track of drift (see step 5). You should also be ready to actively monitor for representation and accuracy by different user segments, such as gender and race, depending on your end-user demographics.

Ensure that you still have human-in-the-loop in production to find instances of bias, react to edge cases, and catch things that were missed by the machine. Structure your processes so that you can expect to receive feedback from your model (e.g., allow for the model to fall-back in cases where it has low confidence, or where it can't make a decision), and also make sure you're able to give feedback to the model in turn. This feedback loop will give your model the flexibility to handle more use cases and reduce erroneous labeling.

When you're confident in the quality of your model, you can move into production.

Testing

Create a controlled production environment to help you track your model's performance. Model performance isn't exclusive to business performance, but also ethical performance. You should have ethical measurements in place to benchmark against. Check if your test cases perform equally across both business and ethical measurements. Do test cases from all your end-user segments perform equally? If not, investigate why.

Think always about the end-user. Ask yourself throughout the build process: am I thinking about my end-user enough? Test your model with a sample of those end-users to find solvable problems before deployment.

Validating and Tuning

Check manually for bias. Have your team review results for both traditional and edge use cases. Is there bias present in the decisions your AI is making?

Correct deficiencies. Take the time to go back and correct any flaws in the data. This may require using other datasets or simulating data to get the results you require.

Validate with new data. Once your model is trained, validate it by using a set of data that it wasn't trained on to get an unbiased estimate of the skill of the final model. Tune the model as needed.

Expect to iterate and learn. Your team should be flexible in adjusting datasets, changing labels, reviewing class types, and overall making changes as needed during the model-building process.

A data partner will assist during validation by testing whether the data was labeled correctly and with the right intent.

Step 5. Actively Learn and Tune

Congrats! You've deployed your model—but how long will it accurately perform in production? Regular retraining is critical to avoid model drift and account for external context transformation. We'll use the example of language changing over time to illustrate different types of model drift:

1. **Sudden drift:** a new concept occurs in a short period of time; e.g., a celebrity coins a new slang term that goes viral.
2. **Gradual drift:** a new concept gradually replaces an old one over a period of time; e.g., the word "basic" has gradually replaced the word "mainstream."
3. **Incremental drift:** An old concept incrementally changes to a new concept over a period of time; e.g., the word "awful" used to mean awe-inspiring, but now refers to something extremely bad.
4. **Recurring concepts drift:** An old concept may recur after some time; e.g., bringing back a slang term like "gnarly" into popular usage.

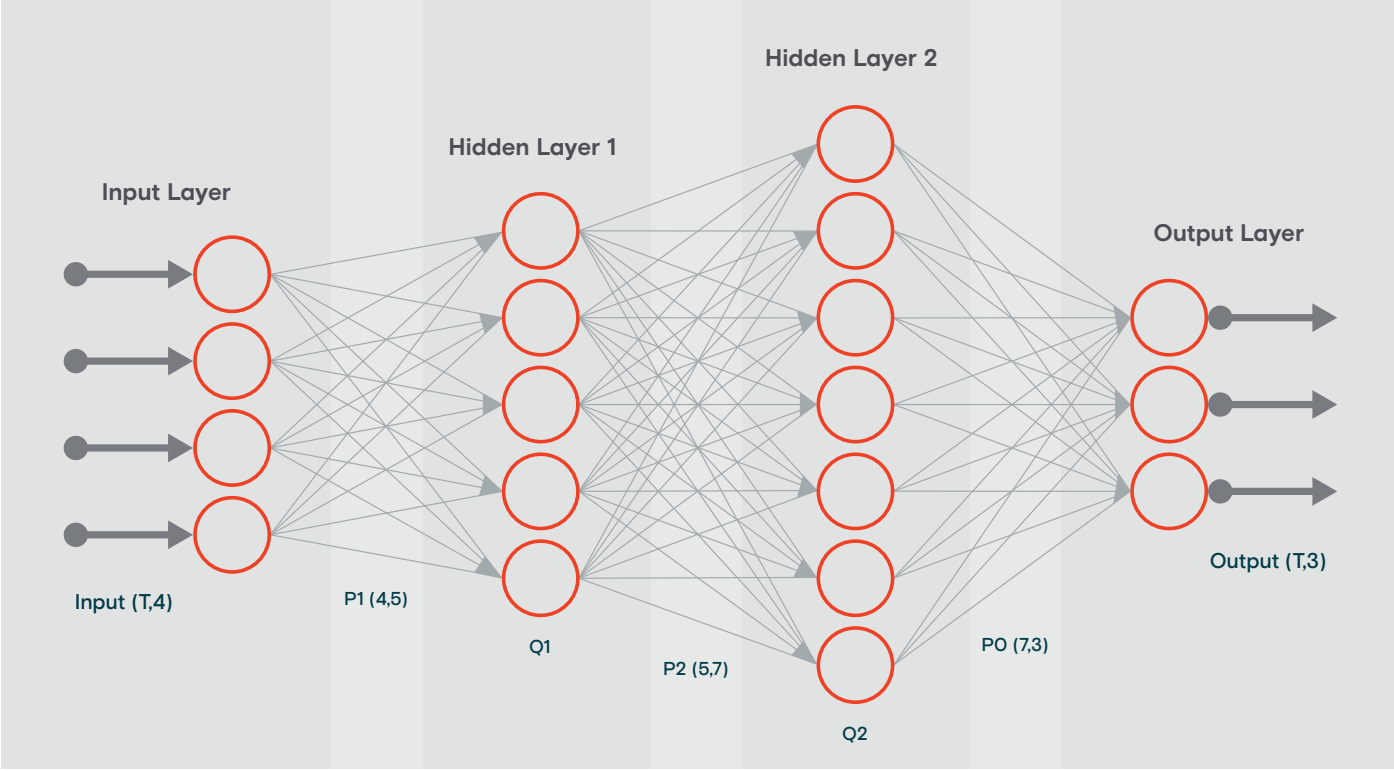
According to McKinsey, one-third of AI products that go live need monthly updates to keep up with changing conditions, like model drift or use case transformation. You can handle drift by building a new model to capture changes or retrain an existing model with new datasets or relabeled datasets. Changing out a model or retraining can be a challenge: pipelines can become complex, and changing one element in the pipeline could impact something downstream. Some models may be feeding into other models as part of the process; changing one will change the results of another. The results could be so subtle that you don't notice. Documentation and practicing explainability is critical to help manage this problem.

Be sure you have an answer to these questions: how can you correct for issues and accommodate change over time? What happens if you find something that was unaccounted for previously? As part of active learning and retraining, you should continue to have humans step in to provide ground truth answers and success monitoring. Be prepared to make changes, and have governance in place on how updates should be made.

Many companies fail in production because they don't retrain their models. As context changes with time, these models will increasingly make inaccurate judgments and become less useful to end-users. Retraining allows you to tune your model consistently, making it more accurate at each point in time. Some successful organizations use active learning to retrain models with new data every day.

The Explainability Problem

While building AI, make a commitment to explainability. Document your build in as much detail as possible, and create clear pipelines through which information is routed.



Why is explainability so important? The cost of training models is climbing as AI becomes increasingly complex and powerful. Some estimates suggest that every year, the cost to make a state-of-the-art model increases ten times that of the previous year.

A current AI model could consist of sub-models and sub-layers with millions of parameters; in other words, so big, you couldn't walk through a single inference step of the model in your lifetime. This is what experts have coined the "black box" problem: inputs go in the box, something happens inside, and the box outputs results. Yet what happens inside the model (or, how a model reaches its decisions, and what methods it uses) can become more ambiguous each training and/or expansion cycle. This has led us to what some call the Reproducibility Crisis.

The need to reproduce models is linked to reducing costs. In theory, it's much easier to re-create or build on what already exists, than start from scratch. Yet, models are now so complex that reproducing them is often an impossible task.

Even in cases where reproducing isn't necessary, models still must be retrained as drift or other changes arise. Without a full understanding of all of the moving parts of a model, retraining can be a challenge and create unintended results (some subtle enough that you won't notice, but your end-users might).

Transparency is also key here. Think about a model that predicts a parolee's likelihood of re-offending. Let's say a judge uses the results of that model to make a judgment on that parolee's future. If you're that parolee, wouldn't you like to know how that model came to that decision? What if the answer is "well, we don't really know"? Explainability can have significant social impacts, especially in regulated domains such as criminal justice, education, or housing.

There's promising work being done on handling explainability. Specific approaches are being explored; for example, holding out certain features or inputs to see how the model works and how it's reasoning. If a model labels an image as containing a guitar, what's the maximum group of pixels that can be left out of that image while still earning that label? Many start-ups are also working on tackling the problem exclusively, but are still a ways out from having a comprehensive solution. In the meantime, the best course of action is to take responsibility and document your work and your data with as much detail and precision as possible, and keep an eye on this space as developments continue.

Looking Ahead

In building responsible, effective AI, remember that there's no trade-off. The steps you take to reduce bias are identical to the steps you should be taking anyway to build a high-quality model. Remember: A biased model that only works for some users and not others is a failed model.

We all have a responsibility to actively control for and mitigate biases to reflect the future we wish to create. Our AI should be representative of our vision. An ethical-first approach to data means AI free from human bias. It's important to recognize that world-class AI has to work for everyone, in every market. Making the effort to reduce bias in AI is paramount so that AI recognizes everything and everyone equally.

If we don't deal with a bias, Gartner predicts that by 2022 85% of algorithms will be erroneous because of it. This can be really bad for business and for the world. This is why all AI teams need to start asking the tough questions, start bringing in stakeholders from all parts of the organization, and start building responsible AI into the core of the business. Now is the time for companies utilizing AI to step up and become good corporate citizens. This means enabling your entire organization — not just the board — to ask the right questions, agree on the same principles, and adopt it as part of the overall culture. Soon, responsible, best practices for AI will not be optional.

It's no longer enough to only think about product design, you have to consider education and training for the people involved in building AI in your organization. Help employees really think about interacting with AI so that they stay focused on deploying technology in an ethical way. You also need to evaluate the entire workflow for AI such as where you are sourcing your data from, if there are proper security measures in place, if the model itself is adding value.

From our work at the World Economic Forum, we found that people from across the world care about the same things, when it comes to responsible AI, and they are very cross-functional:

- Safety and robustness
- Accountability, transparency, and explainability
- Human-machine interaction equity
- Benefits for humans

Move forward, welcome AI into your organization, but do it with the intent of operating responsibly. Look to your board and executive team, as well as to your technical implementors, your product managers, and your customers. Take these initial, cross-functional concerns for ethical practices as the starting place, and work to elevate every aspect of your AI deployments and your organization to be more responsible. **While we're early in implementing responsible AI as a whole, it's up to all of us to keep raising the bar."**

— Kay Firth-Butterfield

Always remember to:

Approach your AI projects with responsibility and inclusiveness in mind. Do the work on the front-end to account for diversity, and there will be fewer surprises on the back-end.

Be empathetic. Understand you will have different end-users and they'll all use your system differently. Consider them throughout your build.

Always be transparent and open about what data you used to train your system, where it was collected, how it was labeled, what the bench-mark for accuracy was, and how it was measured. Transparency enables visibility and explainability.

About Appen

Appen collects and labels images, text, speech, audio, video, and other data used to build and continuously improve the world's most innovative artificial intelligence systems. Our expertise includes having a global crowd of over one million skilled contractors who speak over 180 languages, and the industry's most advanced AI-assisted data annotation platform. Our high-quality training data gives leaders in technology, automotive, financial services, retail, healthcare, and governments the confidence to deploy world-class AI products. Founded in 1996, Appen has customers and offices globally.

- Experience working in **130+ countries**
- Expertise in **180+ languages**
- Over **800 employees** located in offices around the globe
- Access to a curated crowd of over **1 million** flexible contractors worldwide
- More than **13 billion** judgments made and 500,000 hours of audio processed
- **20+ years working** with leading global technology companies

appen