Training, Governance and Accountability (TGA)

Training

The tension in adopting a new technology like AI is the need to encourage experimentation and at the same time the need to manage risk. Experimentation with AI is necessary for people to explore the capabilities and boundaries to develop a formal and intuitive understanding of the capabilities and limitations. Training is necessary so that people become aware of AI's strengths as well as blindspots and weaknesses so that they can use AI responsibly. It is also important to appreciate how use of AI, such as typing in a prompt into ChatGPT or Bard, becomes part of the dataset for the AI (this includes private data added to the prompt). If you haven't done so already, take the 12 question self-assessment to see how well you know the fundamental concepts that power AI.

Training can be engaging and empowering when people get their hands on AI tools and can see the potential for themselves. At the same time, training should reveal the surprising foibles of AI—as well as the more serious faults. In terms of training, we suggest reviewing the fundamentals of how artificial intelligence works by covering pattern-fitting, gradient descent, and compression—and why AI can lack precision and produce biased outputs. Practical case examples relevant to the business or government entity should bring the strengths and weaknesses of artificial intelligence to the fore. Case studies of where AI has made mistakes can help drive home the point AI is fallible. There are excellent hands-on examples of AI bias in publicly available datasets, and step-by-step guides from resources such as the AI Fairness 360 program, developed by IBM and discussed in Chapter 7, as well as resources from Hugging Face, Google, Amazon, and others. There are also specific open-source AI bias detection resources available, and the more technical team

members should be trained on these systems. Safeguards need to be built into the system, and staff need to know how to use them. Business decision makers also need to learn the protocols for model cards and service cards.

In addition to training employees on AI, most organizations need to map out, in a very granular way, the work that is done currently and which tasks AI can replace. Redesigning the entire workflow for AI is often needed. This is the same transformation that software development went through thirty years ago when most computer programs were monolithic structures that required complex and costly rewriting when something needed to change. Many businesses faced a painful realization in the run-up to the year 2000 because their computer systems stored only the last two digits of the four-digit year. Going through these software systems to find and fix the date fields gave many an appreciation for a concept known as *service-oriented architecture* (SOA).

The big idea behind SOA is to break down an application into smaller, independent services that communicate with each other. This allows for more flexible development, testing, and deployment, because changes to one service do not require changes to others. SOA, which began to gain traction in the decade following the year 2000, was transformational for businesses because it enables faster innovation and scaling, improved fault isolation, and the ability to use different technologies for different services based on the specific needs of each service. A *microservices* architecture takes this same approach and extends it to the loosely coupled services which can be developed, deployed, and maintained independently. Each of these services is responsible for discrete tasks and can communicate with other services through simple APIs to solve a larger complex business problem. If an organization maps out the tasks and hand-offs from one person or team to another, it can then determine where an AI can fill in the task. The graphic from Chapter 8 (Figure 8.1) illustrating the change from translating a podcast to another language before and after

the application of AI is a simplified version of how to break down the different elements of a workflow into tasks and hand-offs so that AI can be applied where appropriate.

At the same time, if one transcription service is best in class today but a different vendor has a better synthetic voice, it would be ideal to select a best-of-breed solution for each rather than lock in with a single vendor that is good in one area and mediocre in another. A microservice approach provides flexibility for using different solutions. In addition, given the pace at which new solutions are developed on an ongoing basis, the microservice approach makes it easy to swap out one service for another when something better comes along. Training the organization to see the work that is done as a collection of tasks (services) and hand-offs (similar to application interfaces) will go a long way to preparing the organization for a world where AI is used more extensively to create value.

Privacy and sensitivity of data should also be considered. Few think much about the fact that typing in a search to the Google search engine results in Google saving the search, effectively making it discoverable by a court and visible to Google. When using a large language model (LLM), a person may want to paste large blocks of text or data to get the LLM's response, but one should be mindful that if model is a public rather than private model, the block of text or data essentially becomes public domain content that others may be able to surface in the future. Entering private data into a public AI system could violate privacy laws, or expose sensitive corporate secrets. In addition, when it comes to AI systems, certain architectures of AI can be reverse engineered to reveal the underlying trying dataset.

Governance

A governance framework converts relevant ethical principles to implementable practices in an AI deployment process. In 2022, the government in Singapore developed, in conjunction with ten companies including Amazon, DBS Bank, Google, Meta, Microsoft, Singapore Airlines, Singtel

Group, and others, a model framework for AI governance. Here's an excerpt of Singapore's model framework:

> Guiding Principles:
> 1. Decisions made by AI should be EXPLAINABLE, TRANSPARENT, and FAIR AI systems should be HUMAN-CENTRIC.
>
> From principles to practice, here are the key points in the governance framework:
>
> 1. Internal Governance Structures and Measures
>   - Clear roles and responsibilities in your organization
>   - Standard Operating Procedures (SOPs) to monitor and manage risks
>   - Staff training
> 2. Determining the Level of Human Involvement in AI-augmented Decision-making
>   - Appropriate degree of human involvement
>   - Minimize the risk of harm to individuals
> 3. Operations Management
>   - Minimize bias in data and in the model
>   - Take a risk-based approach to measures such as explainability, robustness and regular tuning
> 4. Stakeholder Interaction and Communication
>   - Make AI policies known to users
>   - Allow users to provide feedback, if possible
>   - Make communications easy to understand

Links to Singapore's model framework, which includes a range of resources for evaluating risk, and other governance considerations, are included in the AI Checklist in the Appendix.

Accountability

Accountability is the active ingredient in an AI governance framework—it is the mechanism that converts the failure to adhere to the values and procedures into consequences. In terms of accountability for removing bias, an organization can be clear on the bias it is trying to remove from artificial intelligence applications by explicitly listing any characteristics that should be safeguarded from bias—race, age, sex, and disability are protected classes for hiring, for example. It would be

extremely problematic for an AI to entrench bias in a range of applications including law enforcement, lending, healthcare, education, and many other categories.

There are two philosophies regarding fairness. One is that the AI model should be tuned to equalize—such as in hiring, where the model should not favor one group more than another. The second philosophy is that the AI should be tuned to achieve the best results possible, and if the results are uneven, then a countermeasure should be applied outside the data and model (for example, partitioning the hiring into two separate models). Those adhering to the second philosophy point out that equalizing the model itself may result in more overall harm as it may require de-tuning prediction or classification to make it worse for some groups—as the system seeks a least-common-denominator. But, the second approach of accepting worse results for some to get the best results overall can only be fair if the governance framework is followed to offset the weaknesses for the groups where AI doesn't perform as well.

The governance framework should specify the operating procedures for transparency when AI is used, the responsibilities for a human in the loop, and how the system will be periodically tested for bias. It should also define the consequences for individuals or vendors that do not follow governance procedures. When someone fails to follow AI governance, is it a fireable offense? Are there clawback provisions in an AI vendor contract for failure to adhere to the governance framework so that there is a financial penalty for the vendor's failure? In other words, what is the accountability in the AI governance system?

RACI is a framework in which people in an organization are classified in relation to decisions they make: those responsible for the decision, those who are accountable, those who require consultation before a decision is made, and those who are informed after a decision is made. For day-to-day operations of AI, a RACI should be established to ensure fair and equitable use of AI. Questions like these should be posed:

- What is the RACI to ensure that the governance framework is implemented and that standard operating procedures are followed?
- Who is consulted before the AI system is launched?
- Who is responsible for ensuring the AI system is not biased?
- Who is responsible to ensure that the use of AI is transparent and ethical?
- After the AI system is launched, are there decisions that require a human in the loop to approve actions?

We are fans of the Japanese quote from quality management: "A defect is a treasure." It expresses the sentiment that we can learn from mistakes and improve the system. Does the accountable person have the authority to dig in to understand the root cause of the mistakes? Does the accountable person have the authority to change the AI system or the governance procedures to address root causes? What authority does the responsible party have for ensuring procedures are followed by others?

Another aspect of accountability is the responsibility to the person whom an AI decision affected. For example, in the law enforcement scenario, what is the responsibility to the person being arrested? Have protocols been put in place to ensure that the AI is correct? What is the recompense for a wrongful arrest, especially when the rules governing the use of AI are not followed? Governance systems work best when paired with accountability.

Organizations should consider working with peers to set-up independent review councils to co-fund independent checks of leading vendors. Or, if the organization has the skill set and financial wherewithal, perform the independent checks of vendors directly periodically and update training accordingly.

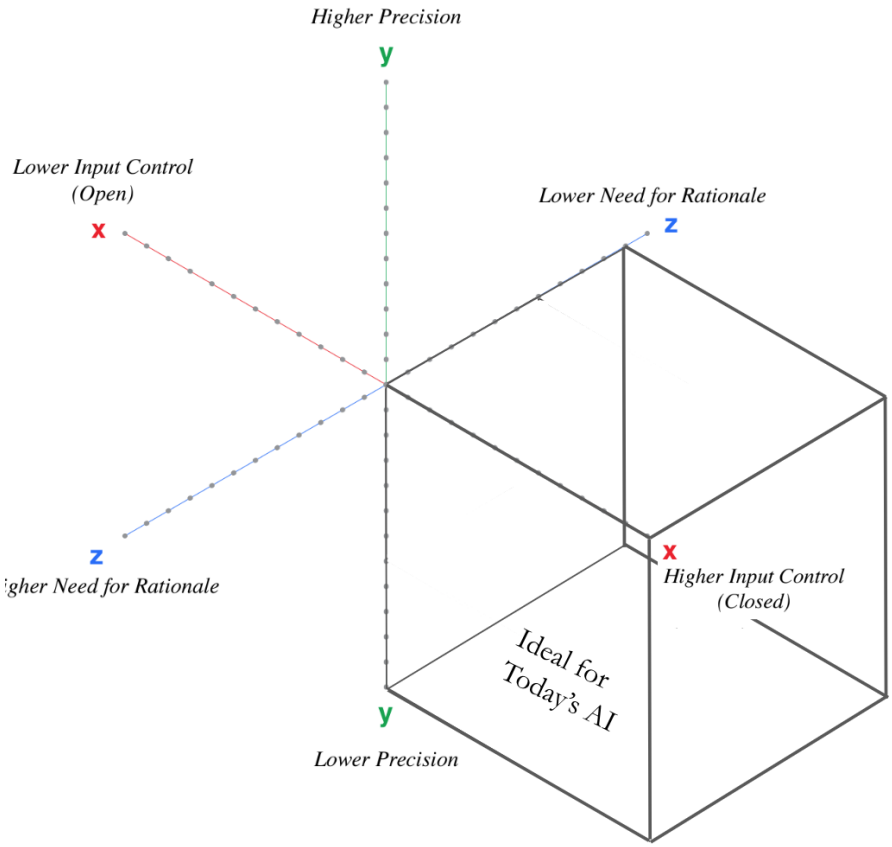
A Checklist for Applying Artificial Intelligence

To aid in the assessment of determining when (and when not) to apply AI and how to apply countermeasures to offset weaknesses, we have created a worksheet of 25 questions. If you are involved in approving AI, use this list as a starting point and add your additional questions. Considering the limitations of today's AI, we would all be well served by asking more questions and contemplating more countermeasures to offset AI's weaknesses

1. What is the current practice for the identified process that will be replaced by AI? (That is, what is the next best alternative to AI that you are currently using?)
    - For the current approach, what is the breakdown of all costs, including financial and time for everyone involved?
    - For comparison, what is the breakdown of all costs for the AI approach?
2. What are the benefits or needs that will be gratified by using AI?
    - Are there cost (direct and indirect) savings for adopting AI?
    - Are there time savings for adopting AI?
    - Is there a performance benefit where AI is more accurate?
    - Are there other benefits for applying AI?
3. What other alternatives, besides AI, exist to improve the current approach, and how do they compare to AI in terms of cost, time, ease of use, and effectiveness?
4. Is the data that AI would use labeled automatically, or will it require building the dataset?

- Are there virtuous cycles in the data feedback loop that will make the AI data better over time?
- Is there risk of vicious cycles with the feedback loop of data influencing the AI?

5. Is the application of AI considered very risky, risky, or not risky based on precision, input control, and rationale? (See Chapter 6 for full details on this framework.)



Lower Precision (y-axis, bottom of chart slice)

		z-axis	
x-axis		**Input: Open (Uncontrolled)**	**Input: Closed (Within AI's Control)**
	Lower Need For Rationale		
	Higher Need For Rationale		

z-axis

6. What are the benefits of the current approach that will be lost if AI is applied?
    - Control?
    - Transparency?
    - Employee satisfaction?
    - Others?
7. What is the truth set for accessing precision of current practice vs. AI?
    - How confident are you in the truth set?
    - Could the truth set be biased?
8. How will you measure precision? Will you use a false accept / false reject method, or some other approach?
    - How precise is current practice? How precise is the AI approach?
    - Will the AI and the next best alternative make the same mistakes or different mistakes?
    - Is there more risk or cost to the type of mistakes AI makes compared to the next best alternative (such as human operator errors)?
    - Today's AI provides an output, such as a reply to a prompt, classification or decision, but not a rationale for the output. Today's AI does not provide an explanation as to how the AI arrived at the output. Is there an alternative to AI that does provide a rationale for the decision?
    - How important is the rationale for the output?
    - In considering the alternative to AI, is there an explanation for when the decision is wrong?

- If someone disagrees with the output of the AI, what is the process of reviewing the decision?
- How will disagreements with the AI output be resolved? And who will make that decision?

9. Are the AI inputs from an open or closed system?
    - How easy or hard would it be for the inputs to be manipulated by a bad actor?
    - Even without a bad actor, is there risk the inputs could evolve with changing industry standards? If so, how can you address this?
10. Over time, how much is AI expected to improve? How much will the next best alternative improve?
    - How will you benchmark to see if AI is on track with the improvement rate?
    - What if AI does not meet the expected rate of improvement?
    - Who will perform benchmarking, and will their measures be objective and informative?
11. Who will perform the risk assessment?
    - What is the worst that can happen if AI gets the output wrong?
    - How serious are the consequences of a wrong output?
12. Will people be in the workflow when AI is adopted?
    - Can humans complement AI to offset weaknesses?
    - In the case of recommendation engines, what role, if any, is there for a human to curate products or otherwise design the user experience to reduce polarization?
    - Does the AI process and reaction time allow for human intervention?

- How will you test for habituation? That is, how will the organization make sure the human in the loop is truly checking the AI and not assuming the AI will be right all the time? At what level of missed human interventions would you stop using AI (or pause AI and re-design countermeasures)?
- How much time and cost does AI + human(s) add to the process? How does the AI + human process compare to the next best alternative in terms of precision?

13. What is the risk of hollowing out a large number of jobs?
    - Are there other benefits to a human in the loop (and lessening the hollowing out problem)?
    - What investment in retraining will the AI owners pursue for displaced workers?
    - What is the strategy for transitioning from human-led activities to AI-led activities? Will humans be engaged in essentially training the AI through a transition period of monitoring and corrective actions? Or, is there a flexible cut-over from human to AI?
    - What is the risk of AI hollowing-out an important skill set that may be difficult to re-acquire if AI isn't as successful as initially expected?
14. What is your back-up plan if it turns out that AI isn't fully up to the task just yet?
15. Is AI expected to access value and trade?
    - Is the data fully labeled and available to the AI, or is there some hidden information that influences value?
    - Does the AI have a unique speed and volume advantage in trading?

- Will the AI have to trade with a human on the other side of the transaction? If so, What is the risk of adversarial selection? Will the cost savings on efficiency be greater than the variance in value assessment accuracy?

  What is the risk of adversarial selection?
- Have you war-gamed how the AI might respond to bad actors seeking to game the AI algorithm?

16. Can your application of AI influence attitudes or behaviors over the longer-term (e.g., personalization AI such as Facebook, YouTube, Amazon)?
    - What are the risks from your content and AI personalization in terms of polarization?
    - If AI is used for brand advertising or product recommendations, how will the advertising content be curated, and algorithm set to ensure a common denominator that contributes to brand value?
    - Can brand journalism be applied to support diversity and inclusivity while preserving brand identity?
17. Can changing the role of AI (the objective function) allow AI to add value without creating the same level of polarization as when AI is optimizing for engagement?
    - Can recommendations be set to pull people together rather than drive them apart?
18. In social media and discussion boards, is it possible to include a credibility or reputation score to improve the quality of content and sharing?
19. How is training data collected, and what is the risk of bias?
    - Are outside training sets or libraries used for any of your AI? Is there any risk of bias in third-party AI contributions?
20. Could adding in a randomized control test to data collection reduce the risk of bias?

- Can the experiment be blinded? Are there other considerations in the design of the experiment that would be necessary to remove bias?

21. To address bias in the AI, can the model be partitioned to avoid AI bias for protected classes or otherwise add value without perpetuating bias?
    - If partitioning is applied, does a quota need to be set, and, if so, who will set it and on what basis?
22. Is fairness testing (such as AI Fairness 360) applied on a regular basis to check for bias?
    - For the full kit of AI Fairness tools, and explainers, visit http://aif360.mybluemix.net/ (open source, supported by IBM)
23. Does your organization have a governance framework that defines your guiding principles and standard operating procedures?
    - There are several model frameworks. Here is one developed by the Singapore government in conjunction with a diverse set of 10 leading businesses: https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework
24. Does your organization have a training module that teaches the team to recognize the hidden weaknesses of AI (lack of precision, risk in open environment, difficult to observe rationale for decisions) and how to offset these weaknesses in your specific applications? If not, do you have a plan to introduce training?
    - Is your governance framework and training paired with transparency and accountability—that is consequences for those that do not follow the governance framework policies?
25. Accountability: Is there true authority among those responsible for living up to the governance framework?

- If the standard operating procedures are not followed, what is the consequence? What are the fireable offenses?
- When a problem is identified, who owns fixing the system so that problem can't happen again? Is there an annual audit?