Tier 4

Tier 4 – Integration: “I want to use the car to make my life easier, my job easier, and understand the rules around using the car safely.”

Tier 4 represents using AI to integrating AI into real workflows. The focus is no longer “what is AI” in the general sense or how to prompt, but how AI fits into systems, processes, and decision-making environments where risk, quality, and accountability matter.

This tier introduces:

- How AI systems operate overtime (the AI lifecycle)
- How humans and AI collaborate in practice
- How AI outputs fail – and why
- How organizations manage AI risk through governance

This tier assumes basic familiarity with AI concepts but does not assume technical expertise.

Overview

At the Tier 4 - Integration level, AI is no longer treated as a novelty or as a chatbot or standalone assistant. Instead, it begins to get treated as a part of a broader system that includes:

- Data
- Humans
- Processes
- Risk Controls
- Accountability Structures

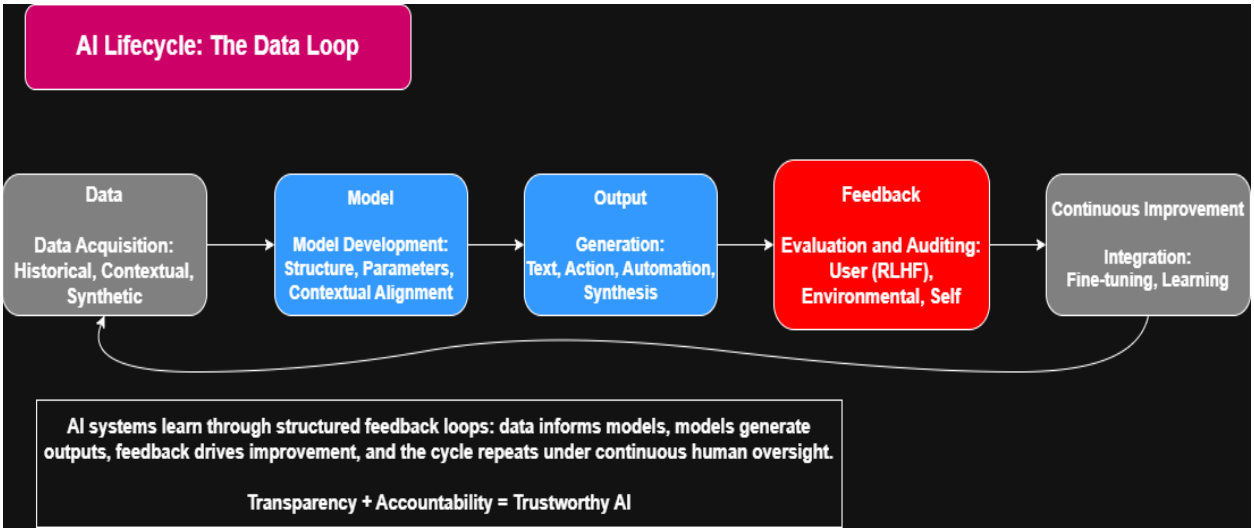
Key focuses of this tier:

- The AI Lifecycle
- Human-in-the-Loop Workflow
- Designing workflows that safely include AI
- Evaluating and validating outputs
- Governance of AI use and results

Note: The concepts of this tier apply primarily to LLM and similar large-scale AI systems, not smaller, single-purpose algorithms.

The AI Lifecycle

Complex AI systems do not operate as static tools: they operate with a continuous lifecycle loop that influences how reliable, aligned, and safe their outputs are over time. The lifecycle continues looping until retirement or the user starts a new session.



Stage 1 – Data

Data is the foundation of any AI system. Inputs can include:

- User prompts and context
- Sensors
- Public Databases
- Internal Databases

Key Activities include:

- Acquiring Data
- Clearing and removing errors
- Labeling and structuring information
- Confirming accuracy and relevance

Poor data will always lead to poor outcomes, no matter how advanced the model is.

Stage 2 – Model Development

This is the process of teaching an AI system how to do something, like training a brand-new employee. The difference is that models are trained specifically using data, instead of experience.

- Engineers define the goal and then design the model architecture around that goal

- The model is trained on specific datasets
- The system learns to identify patterns and relationships within the data
- Stress testing is performed before release to the consumer

Models are not trained to “know” but to predict next likely outputs based on patterns in data.

Stage 3 – Output Generation

The model produces output such as text, classifications, images, and predictions, which appear within the application windows.

- These outputs are constrained by different “layers” to make them safe and reliable, often called “safety guardrails” or just “guardrails”.
- This output is what users are directly interacting with, and what they are able to look at and interpret using their own knowledge and context of the situation.

Stage 4 – Feedback

Outputs of the AI can be evaluated by:

- The users directly
- Internal and external auditors
- Automated monitoring systems (similar to cybersecurity automations)

It is important to look for the following during feedback:

- Errors (such as hallucinations and biases)
- Drift from expected behaviors
- Unexpected or unsafe responses

Results are logged and reviewed.

Note: when interacting directly with LLM, it is important to evaluate outputs and explicitly tell the model when something has both gone right and, more importantly, when something has gone wrong.

Stage 5 – Improvement and Retraining

Feedback and evaluation of outputs are only useful if they are applied to enhance model alignments.

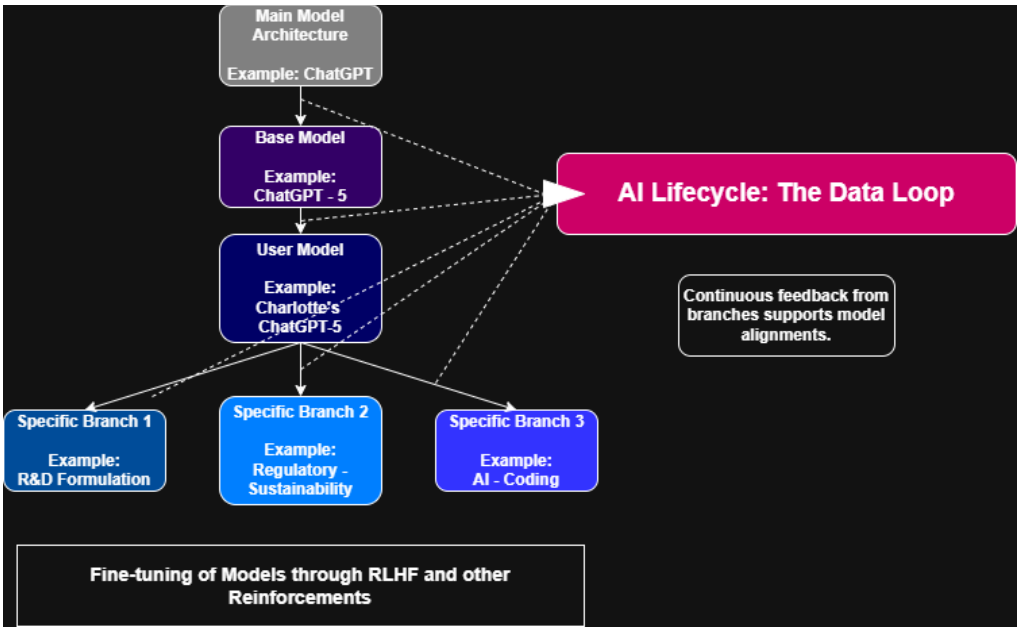
This stage includes:

- Incorporating feedback into new datasets
- Adjusting parameters

- Retraining or finetuning the model

This process is ongoing and is never truly “finished”.

Note: Fine-tuning of the model only happens within context of the window being used (unless the changes are hard coded in).



Human-in-the-Loop (HITL) Workflows

Human-in-the-loop workflows are the most common and safest form of initial AI integration into existing processes. In these systems:

- User gives input.
- AI generates output.
- User reviews, validates, and approves results.
- Human users are always accountable for final decisions.

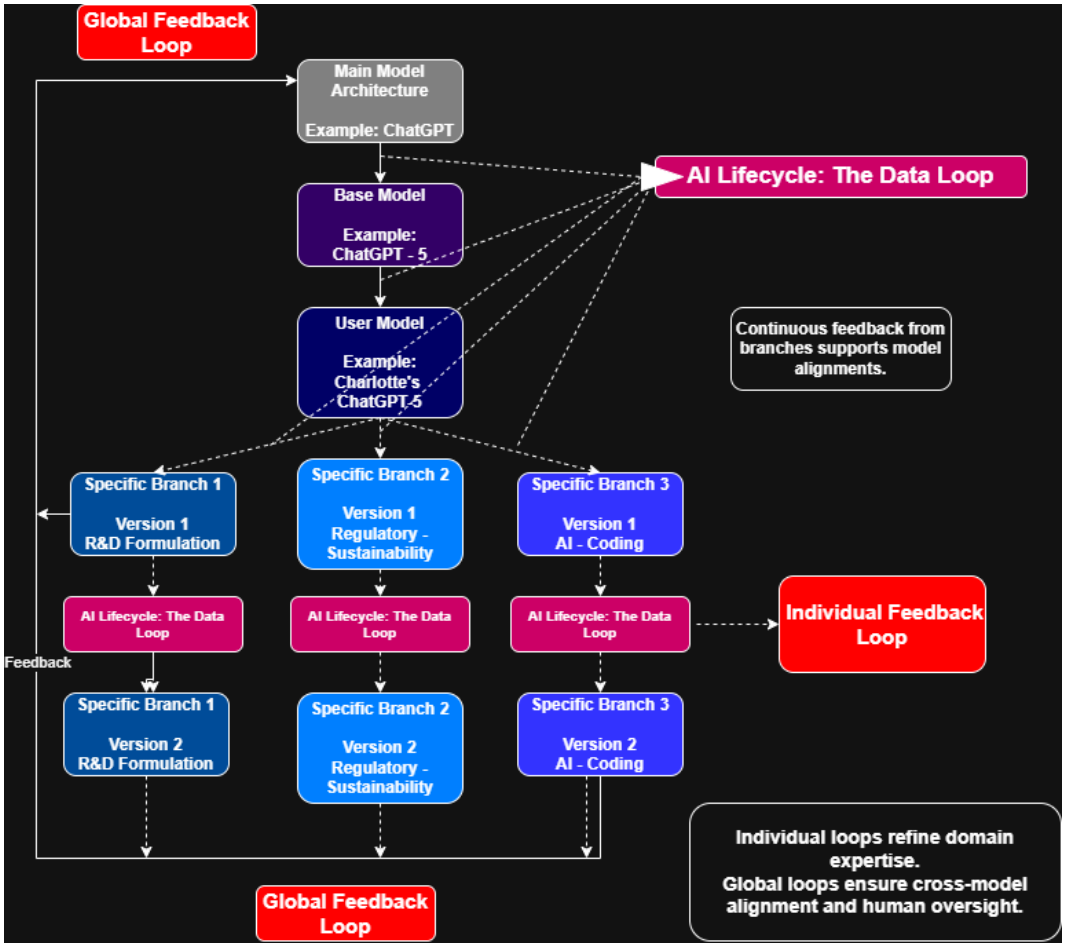
Key characteristics:

- AI assists to speed up the process but does not replace final human decisions.
- Human users act as the final managers of outputs.
- Feedback from users influences future behavior.

Sometimes this is called refinement or fine-tuning; individual fine-tuning is sometimes called “user-optimization”.

Note: There is a current movement towards redefining the word "system" to indicate the intrinsic interplay of humans and AI, not just the AI itself.

- *This signals the move towards treating human-AI collaboration as one interconnected system working in tandem, not two parallel systems working parallel to one another.*



"Don't Be THAT Guy" – Feedback Matters – Real World Example

Every input and piece of feedback into the model reinforces future behaviors on an individual basis, even if only within the specific session or window of interaction.

Strong internal feedback loops have risks for:

- Future repetition of bad outputs
- Rewarding engagement over accuracy
- Training the model to "make the user happy" rather than to inform "factually"

When enough reinforcement accumulates, these behaviors can leak into broader feedback loops. This becomes especially true during:

- New model rollouts
- New model capability shifts
- Reduced moderation periods

Some systems prioritize user satisfaction so heavily that truthfulness degrades and hallucinations increase. ***This is not "malicious misalignment"; it is a predictable outcome of the reinforcement and training design of the model chosen by the parent company.***

Example: The Grok "Mecha Hitler" Episode

Designing Workflows with AI

AI should not be randomly inserted into workflows without proper internal onboarding.

Prior to onboarding an AI augmented workflow, organizations should assess:

1. A definitive goal.
    a. Not a vague idea of "I want to use AI to be better" - helps assess what you need going forward.
2. Initial starting point of infrastructure, financial versus time versus capability limits, and data state of the current system.
    a. Infrastructure – Does the organization have the hardware and software capable of doing what they want to do?
        i. Many organizations fail at implementation because their computers are not made to run the systems they want to use.
    b. Initial costs – What does the organization want to pay in terms of financial investment versus time investment? What capabilities do the employees have right now?
        i. The answer to this question helps to narrow which AI systems to use and how.
        ii. Example: A company can almost always immediately implement a third-party AI platform assistant by the end of the day for a relatively low initial investment, but a customized AI system designed specifically for your industry and current procedures could cost thousands to millions of dollars with months to years until full deployment.

    c. Data state – Is your data organized in a way that a computer will understand it? Is it structured? Is it clean?
        i. Ask yourself if someone knowledgeable in your particular field would be able to understand how to navigate your data by themselves or if it would require your direct input to understand it.
3. Where to add AI value.
    a. Assess the goal and where the process would most benefit from utilizing AI's strengths, such as its pattern matching, its quick reading and summarization capabilities, or its automations of repetitive work.
4. Where human oversight is mandatory.
    a. Ask yourself where you would want a manager to sign off on something versus an everyday employee signing off on something.
5. Where risks are the highest.
    a. Ask yourself what are the worst things that could happen in this process: these are your CCP's.

Adapting the Critical Control Point (CCP) Model

- Borrowed from the widely recognized food manufacturing process, CCP theory asks
    - "Where in the process can something go wrong and introduce risk?"
- In AI workflows:
    - These are the moments where incorrect input could cause real world harm. Examples:
        - An automated purchasing process buying a year's worth of supplies instead of a month.
        - Publication of incorrect information on a broadly released document to clients and consumers.
        - Building a recommendation for a client based off of faulty information pulled from a source which is not considered credible.
    - These are the points that require human verification: ask yourself where you would want a manager or senior employee to check the work of a brand-new hire before action is taken.
- Because every business is different, it is up to each organization to define its own CCP.
- CCP Theory emphasizes PREVENTION and MITIGATION of harm, instead of retroactive correction after the harm has already taken place.

Common AI Output Problems

AI outputs often fail in predictable ways.

1. Hallucinations – Confidently presenting false information
    a. Hallucinations typically occur due to missing information within the datasets the model is trained in, followed by reinforced behavior to keep users happy.
    b. Reinforcement of "user engagement" very often leads to the model filling in information instead of saying "I don't know".
    c. This may be:
        i. Reinforced behavior from previous training cycles,
        ii. Systemic design choices,
        iii. Or both.
2. Bias inherited from training data
    a. Biases within training data becomes reinforced and amplified if it is never corrected.
3. Literal interpretation without context
    a. Models only know what is presented to them; without presenting context, they cannot infer outside constraints.
4. Errors caused by missing information
    a. This missing information often leads to biases and hallucinations.
    b. This is why you often hear that "Data is the new oil".
5. Platform Instability
    a. Ongoing updates and rollouts, especially when there are significant changes to internal structures, often create unreliable outputs.
    b. Metaphor: Taking your normal route to work and realizing there is construction, so you decide to take another route. Then in another week, that route is also under construction and is being rerouted.
    c. It is important to be able to identify these periods as a user to be able to recognize when heightened vigilance for incorrect outputs is needed.
6. Source Choice
    a. Different platforms weight sources differently, sometimes causing divergences in what information is used and presented in outputs.
    b. Example: Gemini is built on Google, so it favors google searches; Grok is built around X so it leans towards X as a source of information.
        i. Note: Many of this source favorability is not "malicious" but is done to speed up the AI and lessen cost per output simultaneously.
7. Mode Collapse due to over-sampling or lack of variability
    a. AI systems rely on math, particularly statistical analysis to pattern-match accurately.

    b. This can lead to overfitting of data (called mode or statistical collapse).
    c. Can be a problem for highly theoretical domains such as Research and Development.

It is important to note that these are systemic behaviors: while they are not driven inherently by user behavior, it is up to the user to spot them.

Mitigating Output Risks

1. Always double check the sources.
    a. Be sure to click on the source links and check to be sure that the link says what the model says it does.
    b. You are checking both that the link exists and that the information in the link is relevant and correct.
2. Use multiple models for cross-verification.
    a. Can be done using the same platform or different platforms.
    b. Utilizing multiple platforms for highly theoretical work is good for catching blind spots due to how difference in how models are trained.
3. Ask the model for "Adversarial Testing".
    a. The model will attempt to poke holes in your output.
4. Using Variability and Verbalized Sampling to mitigate model collapse.
    a. Tell the model to "give me the top five choices".
    b. This forces the model to show you a broader range of the internal outputs before collapse into the "most likely output".
    c. Note that some models, like ChatGPT, have this capability built into main chat interfaces now, but you can force variability by asking for a wider range of possibilities.
5. The more specific you are, the better your outputs will be for your specific needs.
6. Unfortunately, for platform instability, there is not much you can do as a user.
    a. This is tied to the parent company doing the work it is supposed to be doing.
    b. The important part is to be able to catch when it is a period of instability and to adjust workflows accordingly.

Remember: if the task is high stakes with immediate real-world consequences, don't use AI for final outputs; use it for brainstorming and drafting.

Governance

Definitions:

1. Governance: systems of rules, controls, and processes that ensure things are trustworthy, safe, and well-managed.
2. AI Governance: ensures that AI is accurate, fair, explainable, transparent, auditable, and legally compliant.
3. Data Governance: ensures that data is high-quality, secure, and organized

Governance is about structure, not restrictions.

AI systems work best with definitive rules and guidelines, not general ideas of rules. Specific rules and procedures that define what the system can and cannot do, as well as predetermined CCPs at important decision-making boundaries or during abnormal activity, are what catch AI failure modes and prevent them from continuing or proliferating.

AI Policies

These are internal sets of rules, guidelines, and principles that direct behavior, decision-making, and operations to align with goals. Good AI policies define:

- What tools are approved
- For what purpose those tools are approved for
- Who can use the tools for what purpose
- What data is allowed to be used for what purpose
- What tasks AI can do versus what it explicitly cannot do
- What documentation must be required
- How outputs must be reviewed
- Who is accountable for decisions
- How models are approved, monitored, and continuously audited

The same principles for the use of any tool in business are applicable to use of AI.

Accountability: Roles and responsibilities of users and businesses using AI systems, especially in processes that have direct real-world impact.

Auditability: Every AI-assisted process should have a traceable and reviewable decision process.

Transparency: Letting all stakeholders (clients, customers, regulatory bodies) know of the breadth of your AI use, especially when it is in direct communication or direct implication of their relationship with you.

Fairness and Bias: Making sure that the systems that you are using have mitigated unfairness and bias in training data and training methodology, and that processes are in place that monitor any outputs that may fall under this category.

Continuous Monitoring: Having the processes in place to make sure that the AI systems in use are still suitable for their intended purpose.

Documentation: Every process, procedure, and applicable failure is logged and archived for future use.

Different Governance Frameworks

What is an AI Governance Framework?

AI Governance Framework: The rules, roles, and controls that ensure AI systems do what they are intended to do, and, more importantly, are stopped or corrected when doing the things they are not supposed to do.

Many current frameworks are based on the founding concepts in the OECD Principles or the EU AI Act.

The current state of AI Governance shows how governments and organizations think very broadly about responsible AI use across various domains. The examples in the table below represent a mixture of hard laws (enforceable), soft laws (influence enforceable laws), and the standards that many organizations hold themselves to.

Notes: Some businesses have very specific legal obligations for AI use, such as in the financial, civil, and healthcare fields.
For more information on general AI Governance frameworks, visit https://aiqgate.com/current-ai-governance-frameworks/

Framework	Regulatory Body	Geographic Regulatory Area	Oversees	Focus
EU AI Act	European Union	European Union (27 Members)	Sets rules for how AI can be built and used based on risk level	Safety, transparency, protection of citizens

NIST AI RMF	National Institute of Standards and Technology (NIST)	United States (Voluntary; has global influence)	Provides guidance for AI use	AI risks across the AI lifecycle
ISO/IEC 42001	International Organization for Standardization (ISO) & International Electrotechnical Commission (IEC)	International	Provides guidance for AI use	How organizations set up management systems for AI governance
OECD AI Principles	Organisation for Economic Cooperation and Development (OECD)	International	Provides guidance for AI use	Trustworthy, human-centered AI
TRAIGA	State of Texas	Texas, USA	Framework for responsible AI use in government and business	Emphasizes trust, accountability, and public trust

Key Takeaways

- AI systems are not static and they change.
- Human-AI collaborative models need to consider both the human aspects and the AI system characteristics to operate well.
- Knowing how failures can happen and how to spot them enables users to mitigate them from happening in the future.
- Governance frameworks give guidance on using AI systems in a way that benefits humanity.

"The best way to prevent future problems from happening is to learn how to spot them, finding out why they happened, and then using that knowledge to build a better system and process going forward."