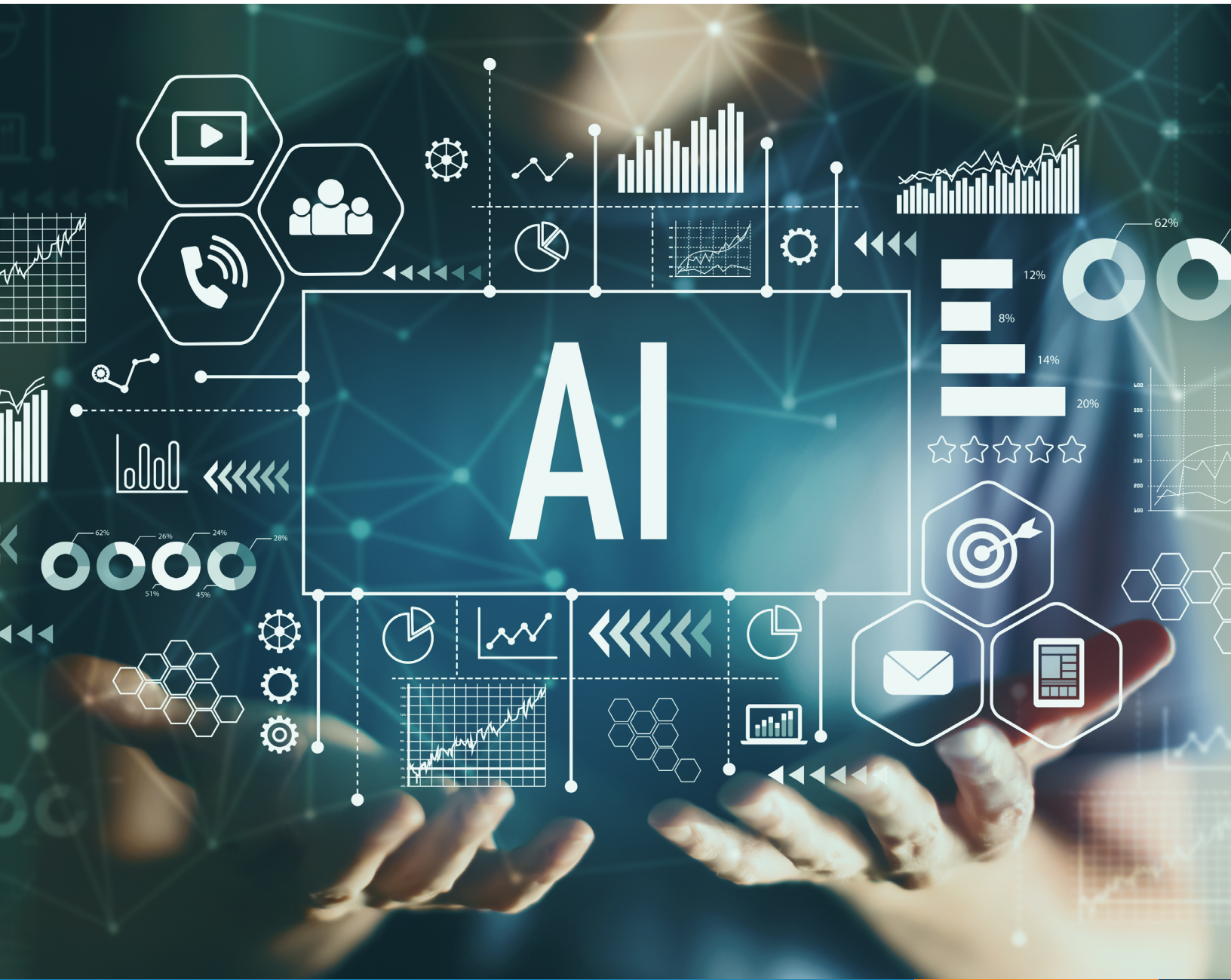
ONLOGIC

AI Crash Course

Everything You Wanted To Know About AI But Were Afraid to Ask ChatGPT



ONLOGIC

Table of Contents

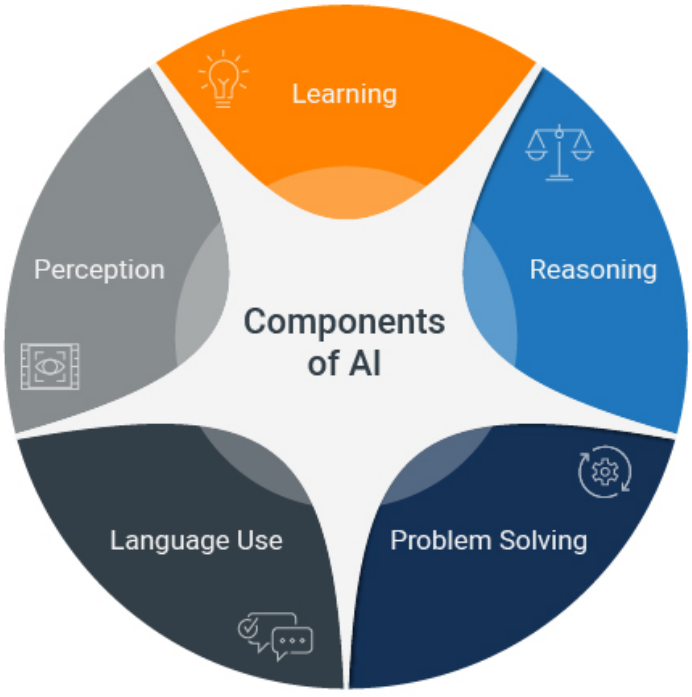
A Crash Course in Artificial Intelligence (AI)

AI has quickly ascended to supreme buzz word status, in no small part due to its potential to revolutionize virtually every industry. If you're still a bit unclear about what AI is, what it can do, or how it might impact your life or your business, you're in the right place.

Read on for a crash course in AI and a look at the elements and capabilities that make it so powerful. We'll also dip our toe into the benefits and opportunities of AI and what you need to do now to prepare to implement AI solutions, including a step-by-step guide. Finally, we've included a glossary of AI terms to help you navigate the AI landscape.

What is AI?

AI, short for Artificial Intelligence, focuses on technologies that attempt to replicate the results or outcomes of human intelligence. To do so, AI must navigate complex information processing including: learning, reasoning, problem solving, language use, and perception.



How does AI work?

AI isn't just a single computer program or application, it is an entire field of computer science. At a high level, there are two phases of AI: training and inferencing.

AI Training

Training for AI is the process of teaching an AI system to identify, process and learn from data. The goal of training the system is to enable inferencing (decision making) once it's deployed. During the training process, large amounts of input data and output expectations are analyzed to formulate a set of structured patterns. These structured patterns are referred to as the 'model'.

The goal is for the AI model to be able to make informed predictions when given new data. Effective AI training requires massive amounts of data and a lot of computational power, often using multi-core processors and GPUs (more on that shortly).

What is a GPU?

GPU stands for Graphics Processing Unit. Sometimes people call them graphics cards or video cards. As the name implies, engineers originally designed GPUs to process graphics including images and videos so that the CPU can focus on other tasks.

GPUs are designed for parallel processing. This enables them to break complex problems into streams of separate tasks and work them out in parallel so they are completed all at once, instead of one-by-one like a CPU.

Turns out, this design is great for more than rendering images and video, the parallel processing capabilities have enabled them to be extremely effective for artificial intelligence (AI), machine learning, computer vision, and more. GPUs eliminate possible processing bottlenecks for faster results, more capabilities, and an improved user experience.

AI Inferencing

AI inferencing is the process of feeding new data to a trained model and turning the data into actionable insights. AI inferencing is where the value of AI is realized, as the models that are built are only as useful as the outputs derived from inferencing.

From a hardware perspective, performing inference operations is generally far less compute intensive when compared to AI training. This opens up the opportunity for AI inference to be performed on lower power/cost processors including NPUs and general purpose CPUs.

Branches of AI

The field of artificial intelligence is large, and getting larger by the day. To break it down, the applications of AI can be divided into several branches, including: Machine Learning, Deep Learning, Natural Language Processing, Expert Systems, Robotics, and Machine Vision.

Machine Learning (ML)

Machine learning models use data and algorithms to perform specific tasks without being explicitly programmed. These machine learning algorithms gradually improve their accuracy over time.

- For example, Plus One Robotics develops machine learning solutions for warehousing and distribution.

In general, there are three broad subcategories of machine learning:

- Supervised machine learning algorithm
  - Data scientists supply labeled training data. Machines use the data to predict outcomes.
- Unsupervised machine learning algorithm
  - Algorithms look for patterns in unlabeled data.
- Reinforcement machine learning
  - The algorithm learns by interacting with its environment and improving behavior based on positive or negative responses to actions.

Deep Learning

Deep learning is a type of machine learning that structures algorithms in layers to create an "artificial neural network". Deep learning is a more advanced approach to machine learning and can be used to solve more complex problems.

- For example, Bear Flag Robotics develops self-driving autonomous tractors powered by deep learning.

Natural Language Processing (NLP)

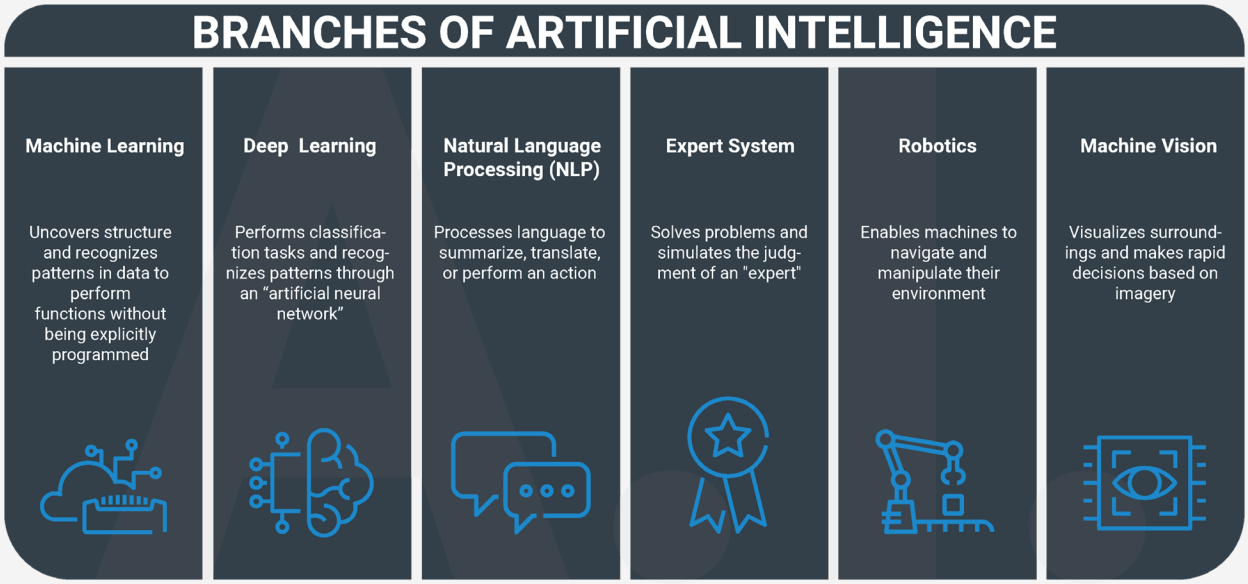
NLP algorithms are used to process written or spoken human language. It is used for translation, summarization, or to perform an action.

- For example, Google translate is an example of NLP to translate text from one language to another.

Expert System

An expert system is software that uses AI to solve problems and simulate the judgment of a human expert.

- MYCIN, a program created in the 1970s by Stanford is an example of an expert system. It is used to aid physicians in the diagnosis and treatment of infections.



Robotics

Robotics is a field of engineering that uses AI to help machines navigate and manipulate their environment.

- For example, advanced.farm creates robotic fruit harvesters that assist in identifying and picking ripe fruits and vegetables.

Machine Vision

Machine vision uses the latest AI technologies to give industrial equipment the ability to visualize its surroundings and make rapid decisions based on what it "sees".

- For example, Artemis Vision creates machine vision solutions for quality inspection in manufacturing to capture minute details that might be missed by the human eye.

Benefits of AI in factory automation

AI offers a host of benefits to factory automation that bring opportunities for reducing costs and producing products of higher quality. Here are our top 6 benefits.

Extended equipment lifespan

- AI-powered predictive maintenance uses sensor data from machinery to detect key attributes like temperature, vibration, pressure and more to identify when it is time to service the equipment. This can minimize unplanned downtime of machinery, reduce maintenance costs, and even extend the lifespan of machinery.

Reduced scrap

- Multiple points of quality control can be incorporated within a production line that can capture quality issues that may be missed by the human eye, dramatically increasing inspection accuracy. If defects are caught early in the production process, scrap can be reduced.

Safer working environments

- AI powered solutions such as quality control can be placed in locations that may not be safe for humans, such as right on the production line.
- Robotic solutions can be created to perform the most difficult and/or repetitive tasks on a production floor.

Enhanced productivity

- AI-driven automation, for example using robotics, can enable work around the clock to perform tasks, even repetitive tasks, at peak performance with a high degree of accuracy.

Improved management of inventory & demand forecasting

- AI can identify trends in inventory and demand to optimize inventory levels, predict demand and even track shipments across a complex global supply chain. It can also ensure that manufacturing decisions are based on the most recent and up-to-date data.

Streamlined factory layouts

- AI can identify inefficiencies in production workflow, clear bottlenecks, and boost output on a factory floor. One way to do this is by creating a factory floor digital twin. Adjustments can be made on the twin to identify outcomes.

AI at the edge

To enable near real-time decision making, many businesses are moving AI solutions away from the cloud and to the edge – nearer the source of the systems creating the data to deliver real-time actionable insights.

AI model building and inference can, of course, take place in a centralized cloud or offsite data centers. However, putting AI capabilities at the network's edge helps to speed decisions and actions based on the information gathered from on-site sensors. The edge of the network might be in a warehouse, on a manufacturing line, on a forklift, or even at a remote, off-site location.

Edge AI example

Autonomous vehicles such as the tractors created by Bear Flag Robotics are an example of AI at the edge. Information from a myriad of inputs including 3D cameras, LiDAR and position sensors is collected and processed to inform vehicles about where to turn to avoid uneven soil, how to adjust equipment, and when to brake to avoid obstacles.

These are all decisions that need to happen quickly – there isn't time to send the data to the cloud, process it, and then return instructions to the tractor. Having the data ingested and processed right on the computer mounted on the tractor rather than in the cloud, drastically reduces that delay, called latency.

Not to mention, access to an internet connection might be a challenge on a farm field. With AI at the edge, the computing can just continue, regardless of where that data is being produced.

Technology for artificial intelligence solutions

The explosive growth of, and resulting value from, AI is closely tied to the explosion of available data, open source tools, and the advancement of technology to process and act on the information gathered. A few of the key advancements include:

Data availability, organization, and classifications

- The Internet of Things (IoT) provides the ability to capture data from a wide variety of connected devices and open datasets.
- Data-centric AI is an emerging science that works towards improving datasets for better performance in practical ML applications.

What is the Internet of Things (IoT)?

The IoT refers to the collection of devices that communicate with each other via the internet. Common IoT hardware devices include wearable devices, smart home devices, and cloud-connected industrial sensors.

What is the AIoT?

The AIoT combines AI with IoT to make connected devices capable of processing and learning information. This, in turn, can help to predict future patterns or events that enable decisions on how to react.

What is open source?

The term open source refers to something that people can modify and share because its design is publicly accessible. For example, open source software is software with source code that anyone can inspect, modify, and enhance.

Advantages of open source AI tools include transparency, collaborative participation, rapid prototyping, and community-oriented development. It can lead to faster progress as developers learn from each other and build upon each other's work. Some also see it as a way to identify sources of bias that may develop in AI and also avoid a corporate AI monopoly.

Availability of open Source AI tools

- Open source AI frameworks, toolkits and pre-trained models have enabled quick AI adoption and provide the building blocks to architect, train, validate, and deploy AI solutions. By eliminating the need for manual coding and traditional software development methodologies, these toolkits can significantly reduce costs and decrease time to deployment.
    - Some examples of toolkits include OpenVino, TensorFlow, Apache, Hugging Face, and PyTorch, to name a few.

Processor advancements

- Multi-core CPUs with integrated GPUs and NPUs (Neural Processing Units), as well as advanced discrete accelerators and GPUs with new AI specific enhancements are foundational to AI's rising value. They deliver the computing power to process and interpret datasets to build an AI algorithm.

> ### What is a CPU?
>
> CPU stands for Central Processing Unit and you will see it commonly referred to as the brain of the computer. Put simply, the CPU is electronic circuitry that controls data – including input, output and storage. The CPU makes sure that the computer operating system and applications all work.

GPU vs CPU

We've outlined the definitions of a GPU and CPU, but what's the difference? Both the GPU and the CPU are silicon-based microprocessors that handle data. Their differences lie with how they process data and their physical processing units – their cores.

Data processing

- CPUs perform serial computing – processing data one task after another.
- GPUs perform parallel computing – processing many calculations at once.

The cores

- CPUs have a handful of cores.
    - For example, Intel® 13th generation Hybrid Core processors are available with up to 24 cores.
    - Note that Intel introduced Hybrid Core technology with 12th Gen. You can read more about that here.
- GPUs might have 100s to 1000s of cores that are similar to CPU cores only smaller.
    - For example, Nvidia's RTX A4000 GPU offers over 6,300 cores including: 48 second-generation RT Cores, 192 third-generation Tensor Cores, and 6,144 CUDA cores.

What is a GPU used for?

Both CPUs and GPUs are really good at math. But as mentioned above, CPUs process the equations one task after another. On the other hand, a GPU can process the equations all at the same time. Because of this, GPUs offer extraordinary computational power for complex mathematical calculations.

Rendering images and graphics boils down to a series of math problems - notably operations of multiplication and addition in a structured manner called matrix multiplication. This supports the different functional blocks of AI. And, as we mentioned, GPUs excel at that. In fact, they accelerate those workloads for superior graphics performance – which is why they are called “accelerators”.

While GPUs were originally designed to accelerate graphics processing, engineers soon discovered that GPUs could handle the advanced challenges of artificial intelligence including the ability to take in a lot of data, classify it and use this information to make inferences and predictions for future outcomes.

Types of AI accelerators

Integrated GPU (iGPU)

An integrated GPU, sometimes called iGPU, is built directly into the computer’s processor or CPU. It shares power and system memory with the CPU. As a result, computers with integrated GPUs are power efficient and generally lighter and smaller. However, because the memory is shared, heavy GPU processing tasks may create latency with general CPU processing.

Discrete GPU

A discrete, or dedicated GPU is a graphics processor that is completely separate from the CPU and has its own memory called video memory or VRAM. With its own memory, it doesn’t have to rely on (some might say hog) the computer’s RAM. Discrete graphics cards deliver better performance but use more power and as a result, generate a lot of heat.

Keeping a discrete GPU cool is a big design consideration in order to achieve maximum performance. Because of this, engineers have designed these GPUs with their own cooling fans which can add to their bulk and energy consumption.

Explore our Beginners Guide to GPUs for more.

TPU

A Tensor Processing Unit (TPU) is an application-specific integrated circuit (ASIC) developed by Google. The primary task for TPUs is mathematical matrix processing which accelerates machine learning workloads. In the words of Google, “TPUs can’t run word processors, control rocket engines, or execute bank transactions, but they can handle massive matrix operations used in neural networks at fast speeds.”

NPU

A Neural Processing Unit (NPU) is purpose-built to accelerate the execution of neural networks for deep learning. NPUs deliver high performance while minimizing power consumption. As a result NPUs are often used in mobile devices, edge computing, and other energy-sensitive applications.

Industrial computers and AI

Industrial computers can be equipped with accelerators such as GPUs for AI and other deep learning applications. Engineers have thoughtfully designed these computers to withstand extreme environments and have enabled businesses to apply AI just about anywhere it is desired.

For example they are used to power robots in manufacturing facilities, warehouses and even in farm fields where they might be exposed to vibrations and extreme temperatures. They are also used to power quality inspection, sometimes right on a production line where the environment might not be safe for humans, let alone a traditional desktop computer.

Considerations for a GPU in an industrial computer

Industrial computers are engineered to withstand extreme environments. A common design feature for industrial computers is their fanless and ventless design. While it offers many benefits, this design doesn't work well for all GPUs since some have significantly higher power consumption, and therefore heat that needs to be dissipated, relative to CPUs.

iGPU for industrial AI

Using and cooling integrated graphics cards that can power a simple image and object recognition solution is entirely possible with a fanless design. Especially when done with some of the newest and incredibly powerful processors.

Discrete GPUs are not always needed and they can add considerably to the cost of a computer. Running AI inferencing on a CPU or integrated GPU (iGPU) helps reduce the cost of effective AI deployment.

When it comes to industrial computers for AI, a fanless design enables a small form factor, which offers a little more flexibility when it comes to installation. In addition, since the computer doesn't require a fan to bring air inside the chassis, these fanless computers can be installed in areas that might have a lot of dust or debris in the air and/or an area that might be subject to vibration.

Discrete GPU for industrial AI

A more complex AI or ML solution may require a discrete or dedicated graphics card. As mentioned above, GPUs generate a lot of heat. That means keeping a GPU cool is an important consideration, and most discrete GPUs come configured with their own fans as well as environmental temperature limitations.

One option is a fanless hybrid computer that provides an isolated expansion bay. A GPU can be installed in this bay and actively cooled with a fan. Meanwhile, the motherboard and other sensitive internal components are protected in the ventless portion of the chassis. In the event you need to replace the GPU, it is easily accessible.

Getting Started with an AI Solution - a Step by Step Guide

So, you've determined that implementing AI might benefit your business and you're ready to dive in. But how do you get started? It can be easy to jump right to hardware and software research, but there are a few vital steps to consider first.

1. **Conduct a Needs Assessment**
    - Identify the areas of your business that might benefit from AI-powered automation. Some common examples include:
        - Repetitive tasks
        - Quality control
        - Predictive maintenance
        - Supply chain management.
    - Through your assessment you can then determine if an AI powered solution is right for your business.
2. **Define Clear Objectives**
    - Establish SMART (specific, measurable, achievable, relevant and time bound) goals for your AI-powered automation deployment. Whether it's improving efficiency, reducing costs, or enhancing product quality, having well-defined objectives will guide your AI implementation and help you assess what's working and what may not be.
3. **Build Your Team**
    - Assemble a team that includes representatives from across your organization - especially operations and IT. Collaborative efforts ensure that the implementation aligns with both technical requirements and production needs.

4. **Assess Your Data Infrastructure**
    - Evaluate the availability and quality of your data. AI relies on large datasets for training models. Start collecting data from all areas of your manufacturing processes.
5. **Collaborate with Vendors:**
    - Engage with vendors and consultants who specialize in automation. They can provide insights, guidance, and possibly custom solutions tailored to your specific needs. Every AI implementation is different, so it's important to work with vendors who are able to commit the time and resources to understanding your unique challenges.
6. **Design your continuous improvement strategy**
    - You'll want to assess:
        - Is the AI model doing its job well enough?
        - Is it starting to miss things?
        - Does it need to be re-trained with new data?
        - How will you push out new models over time?

Hardware for artificial intelligence solutions

When it comes to AI hardware options, different applications will require different AI solutions.

IoT gateways

Whether your AI solution lives in the cloud or an on-premise server, you need an IoT gateway. This small but reliable computer is the go-between for data collected by embedded sensors and the cloud. They serve an increasingly vital role in AI solutions for gathering, storing and sometimes partially processing incoming data before it's transmitted.

For example, we engineered the Karbon 410 for reliability in even the most challenging installation conditions, which might include extreme temperatures and vibration prone locations.

AI at the Edge with an integrated GPU

Some of the newest processors with their integrated GPU can easily support AI inference solutions at the edge. For example, the fanless OnLogic Helix 511 is powered by Intel 12th Generation processors with hybrid core architecture and DDR5 memory. This compact powerhouse offers plentiful I/O including legacy connections and powerful processing for an AI at the edge solution.

AI at the edge with discrete GPU

If you are looking to implement a more robust solution powered by 12th or 13th Gen Intel Core™ processing with discrete GPU, the Karbon 804, a fanless hybrid PC, offers high performance computing and has options for GPU expansion cards with PCIe Gen 4.0 capabilities. The incredible throughput offered by PCIe Gen4 means you have options when it comes to deciding the slots and lanes to use for your GPU.

- Option 1: A single PCIe riser with x16 slots and x16 lanes giving you a throughput of 256GT/s.
- Option 2: A dual riser that offers 2 x16 slots with x8 lanes each. You could use one slot for the GPU giving you a throughput of 128GT/s. For many AI applications, that throughput is more than enough. In addition, the computer will then have an open slot giving you configuration flexibility for additional expansion options.

Need a smaller alternative?

Our Karbon 803 offers a slightly smaller form factor with a single PCIe Gen 4 x16 slot.

GPU Server

For complex workloads, deep learning at the edge, and on-premise training and inference, a GPU server like the AC101 is a great solution. This platform offers Intel 13th generation processors and advanced GPUs and DDR5 memory. Many businesses are using edge servers in their cloud repatriation strategies to move compute resources on-premise to avoid latency and reduce operational costs.

Key Takeaways

As our examples have demonstrated, AI is transforming nearly every industry and changing how we live, work, and play. Now is the time to start strengthening your technology infrastructure to be better positioned to reap the benefits of AI including improvements in productivity, efficiency, quality, and customer satisfaction.

It all starts with the following 5 steps:

1. Conduct a Needs Assessment
2. Define Clear Objectives
3. Build Your Team
4. Assess Your Data Infrastructure
5. Collaborate with Vendors
6. Monitor and improve

From there you can pilot an AI project and scale incrementally. Start small and build on your success and continuously improve your AI models and processes. The OnLogic team is ready to help you understand your hardware needs and how an industrial computer can address your challenging requirements. The future is now and using AI is no longer a matter of “if” but “when”. We’re here to help you get started.

Glossary of AI terms

Algorithm	Software that includes a set of instructions or rules followed by a computer to perform a specific task or solve a particular problem.
Algorithm Bias	Unintentional and systematic errors in an algorithm that result in unfair or discriminatory outcomes, often reflecting existing biases in the training data.
Algorithmic Transparency	The degree to which the inner workings of an algorithm are open and understandable, especially in applications where decisions impact individuals or society.
Artificial Intelligence (AI)	The simulation of human intelligence processes by machines, especially computer systems, to perform tasks that typically require human intelligence.
Artificial Intelligence of Things (AIoT)	The AIoT combines AI with IoT to make connected devices capable of processing and learning information.
Big Data	Extremely large and complex datasets that traditional data processing tools are inadequate to handle, often requiring advanced analytics and machine learning.
Chatbot	A computer program designed to simulate conversation with human users, especially over the internet.
Cloud Computing	The delivery of computing services, including storage, processing power, and AI capabilities, over the internet.
Computer Vision	The field of AI that enables machines to interpret and make decisions based on visual data, often involving image and video analysis.
Data Science	The interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract insights and knowledge from structured and unstructured data.
Deep Learning	A type of machine learning that involves neural networks with many layers (deep neural networks), allowing the system to learn and make decisions on its own.
Digital Twin	A Digital Twin is a virtual version of a physical object, process, or location that serves as a real-time digital counterpart.
Edge Computing	Processing data near the source of data generation (e.g., IoT devices) rather than relying on a centralized cloud server.
Expert System	An expert system is software that uses AI to solve problems and simulate the judgment of a human expert.
Generative AI	Generative AI, also called “GenAI”, is used to create new content (such as text, audio, video, imagery, and even code) based on training data.

GPU	GPU (Graphics Processing Unit) is a computer component originally designed for rendering images and graphics. The parallel processing capabilties of a GPU enable them to break down complex problems into streams of separate tasks and work them out in parallel so they are completed all at once. While GPUs were originally designed to accelerate graphics processing, engineers soon discovered that GPUs could handle the mathematical challenges of many other tasks including AI.
Human-in-the-loop machine learning	Human-in-the-loop machine learning, sometimes called HITL, is a process whereby humans provide input to a machine learning model to maximize accuracy and increase efficiency for applications that use artificial intelligence (AI).
Industrial Computer	An Industrial PC is a robust computer designed to be used in an industrial environment, often for applications like the manufacturing of goods, building automation, and logistics centers. An industrial computer typically has the following characteristics: fanless and ventless design, ability to withstand harsh environments, highly configurable, extensive I/O options, and a long lifecycle.
Inferencing	AI inferencing is the process of using a trained model to make predictions and turn the data into actionable insights.
Internet of Things (IoT)	The network of interconnected devices that communicate and exchange data, contributing to the collection of vast amounts of real-time information for AI applications.
Machine Learning (ML)	A subset of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
Machine Vision	Machine vision encompasses the technology, methods, software, and hardware involved in visual input processing,
Narrow AI (ANI)	Narrow AI is only capable of performing a specific, singular or focused task.
Natural Language Processing (NLP)	A field of AI that focuses on the interaction between computers and humans through natural language, enabling machines to understand, interpret, and generate human language.
Neural Network	A computational model inspired by the structure and functioning of the human brain, consisting of interconnected nodes (neurons) organized in layers.
Neural Processing Unit (NPU)	A NPU is a specialized processor explicitly designed for accelerating machine learning algorithms.
Open Source	The term open source refers to something that people can modify and share because its design is publicly accessible. For example, open source software is software with source code that anyone can inspect, modify, and enhance.

Predictive AI	Predictive AI forms predictions about future events, behaviors, and trends based on historical data.
Predictive Maintenance	A predictive maintenance model uses data from a multitude of sources to predict when equipment is likely to require maintenance.
Prescriptive Maintenance	Prescriptive maintenance takes predictive maintenance a step further. In addition to using data to predict failures, it also prescribes the most effective maintenance actions.
Reinforcement Learning	A type of machine learning where an agent learns to make decisions by interacting with its environment, receiving feedback in the form of rewards or penalties.
Robotics	Robotics is a field of engineering that uses AI to help machines navigate and manipulate their environment.
Rugged computer	Both industrial and rugged computers provide maximum reliability and high performance to run powerful software and control complex applications. (See definition of industrial computer.) Rugged computers have additional design features to make them ultra reliable and operable in particularly harsh environments.
Supervised Learning	A type of machine learning where the algorithm is trained on a labeled dataset, with input-output pairs to learn the mapping function.
Tensor Processing Unit (TPU)	An application-specific integrated circuit (ASIC) developed by Google as an AI accelerator. The primary task for TPUs is mathematical matrix processing which accelerates machine learning workloads.
Training Data	The dataset used to train an AI model, providing examples and patterns for the system to learn from.
Transfer Learning	A machine learning technique where a model trained on one task is adapted for a different but related task, leveraging knowledge gained from the original training.
Unsupervised Learning	A type of machine learning where the algorithm is given unlabeled data and must find patterns and relationships on its own.