Generative AI

Text Generation

Generating text is the task of generating new text given another text. These models can, for example, fill in incomplete text or paraphrase. This task covers guides on both text-generation and text-to-text generation models. Popular large language models that are used for chats or following instructions are also covered in this task.

A model trained for text generation can be later adapted to follow instructions. One of the most used open-source models for instruction is OpenAssistant.

A Text Generation model, also known as a causal language model, can be trained on code from scratch to help the programmers in their repetitive coding tasks. One of the most popular open-source models for code generation is StarCoder, which can generate code in 80+ languages.

A popular variant of Text Generation models predicts the next word given a bunch of words. Word by word a longer text is formed that results in for example:

- Given an incomplete sentence, complete it.
- Continue a story given the first sentences.
- Provided a code description, generate the code.

The most popular models for this task are GPT-based models. These models are trained on data that has no labels, so you just need plain text to train your own model. You can train text generation models to generate a wide variety of documents, from code to stories.

Chatbots

A Chatbot is software created by artificial intelligence designed to interact with people in a natural language. A Chatbot is probably one of the best applications of automatic natural language processing. A *chatbot* is a computer program that can hold a conversation with a human using voice commands, text conversations, or both. Chatbot, also known as a chatterbot, is an artificial intelligence product that can be integrated and used through any messaging application. Chatbots can be divided into two basic types, firstly a Rule-based Chatbot and secondly a Self-learning Chatbot according to the way, how an answer is generated.

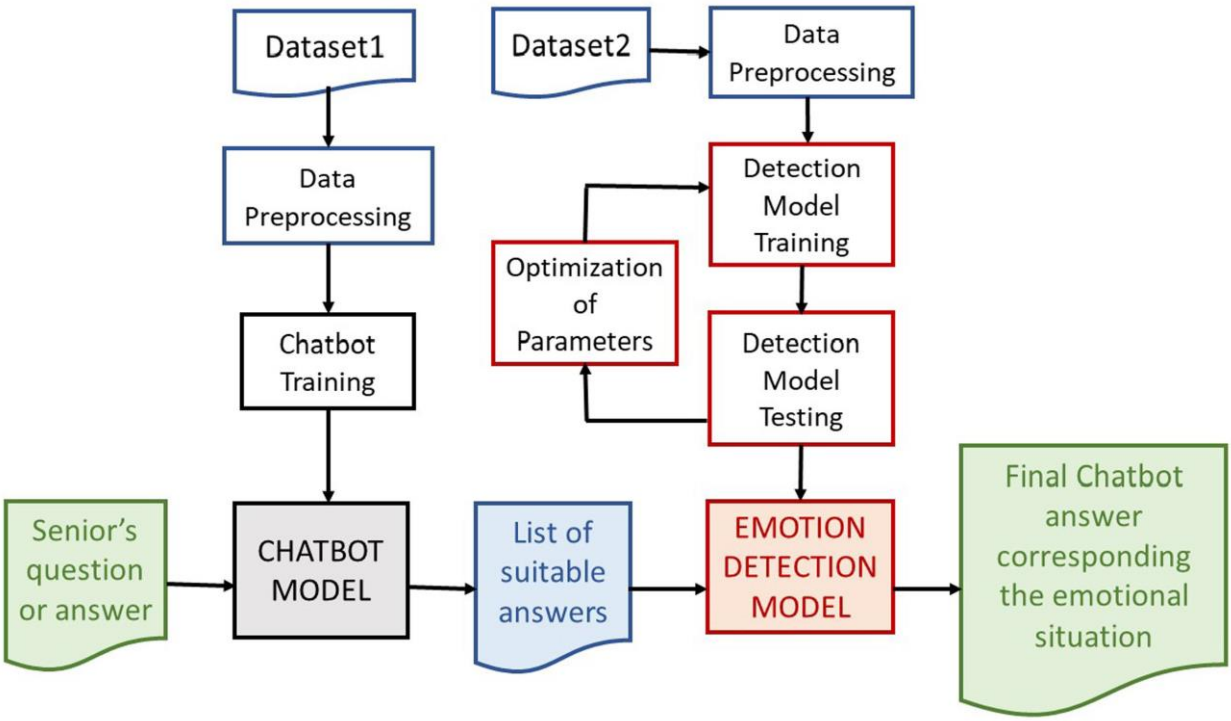
The *Rule-based Chatbot* can answer questions based on a set of predetermined rules that it was trained on. The rules can be very simple but also very complex. This Chatbot is quite good at handling simple queries, but it is not sufficiently accurate in the case of more complex requests.

Self-learning Chatbots use advanced artificial intelligence technologies like machine learning to train themselves from examples. Self-learning Chatbots can be further divided into two

categories: Chatbots learned by loading and generative Chatbots. A *Load-based Chatbot* operates on predetermined input patterns. The Chatbot uses a heuristic approach to provide an appropriate response. It is a goal oriented Chatbot with customized features such as conversation flow. The Chatbot uses a special tone to enhance the customer experience. The *Generative Chatbots* do not use predefined answers. They use the neural networks, "seq2seq" approach when inputs (queries) are transformed into outputs (responses).

Today we have advanced intelligent *AI-powered chatbots* that use natural language processing (NLP) to understand human commands in text or voice forms and they can learn from experience. Chatbots have become an essential customer interaction tool for companies that are active on the Internet. Python-powered Chatbots are handy tools as they facilitate messaging between the company and the customer. Apple Siri, Amazon Alexa, and Microsoft Cortana are worth mentioning. Because these Chatbots can learn from behavior and experience, they can respond to a wide range of queries and commands. The most successful advanced Chatbot today is ChatGPT, which can generate extensive responses in a wide range of domains. This chatbot is a big concern for educational institutions.

We have designed the approach to selection the most appropriate response by chatbot in communication between human and chatbot after considering an emotional state of the person with the help of an emotion detection model. The communication and workflow between a human, the chatbot model and the emotion detection model is illustrated below.



Sentiment Analysis

Sentiment analysis is a scientific field that examines and analyzes the subjective content of textual data from the conversational content of social networks. It mainly focuses on the analysis of the polarity of opinions, attitudes, and emotions of people to determine their satisfaction or dissatisfaction with the object of discussion. It is a challenge for projects in the field of natural language processing, computational linguistics, and text mining. Sentiment analysis includes the following sub-problems.

- Subjectivity detection aims to distinguish subjective from neutral terms, phrases, sentences or comments and is frequently used as an initial step in polarity and intensity of polarity recognition, to separate subjective information from objective ones.

- Polarity classification attempts to classify texts into positive, negative, or neutral classes. It forms the basis for determining the polarity of the text as a whole.

- Intensity classification goes a step further and attempts to identify the different degrees of positivity and negativity, e.g., strongly negative, negative, fair, positive, and strongly positive. Also, special words “intensifiers” are used for intensity classification. They can increase or decrease the intensity of polarity of connected words, e.g., surprisingly good, highly qualitative.

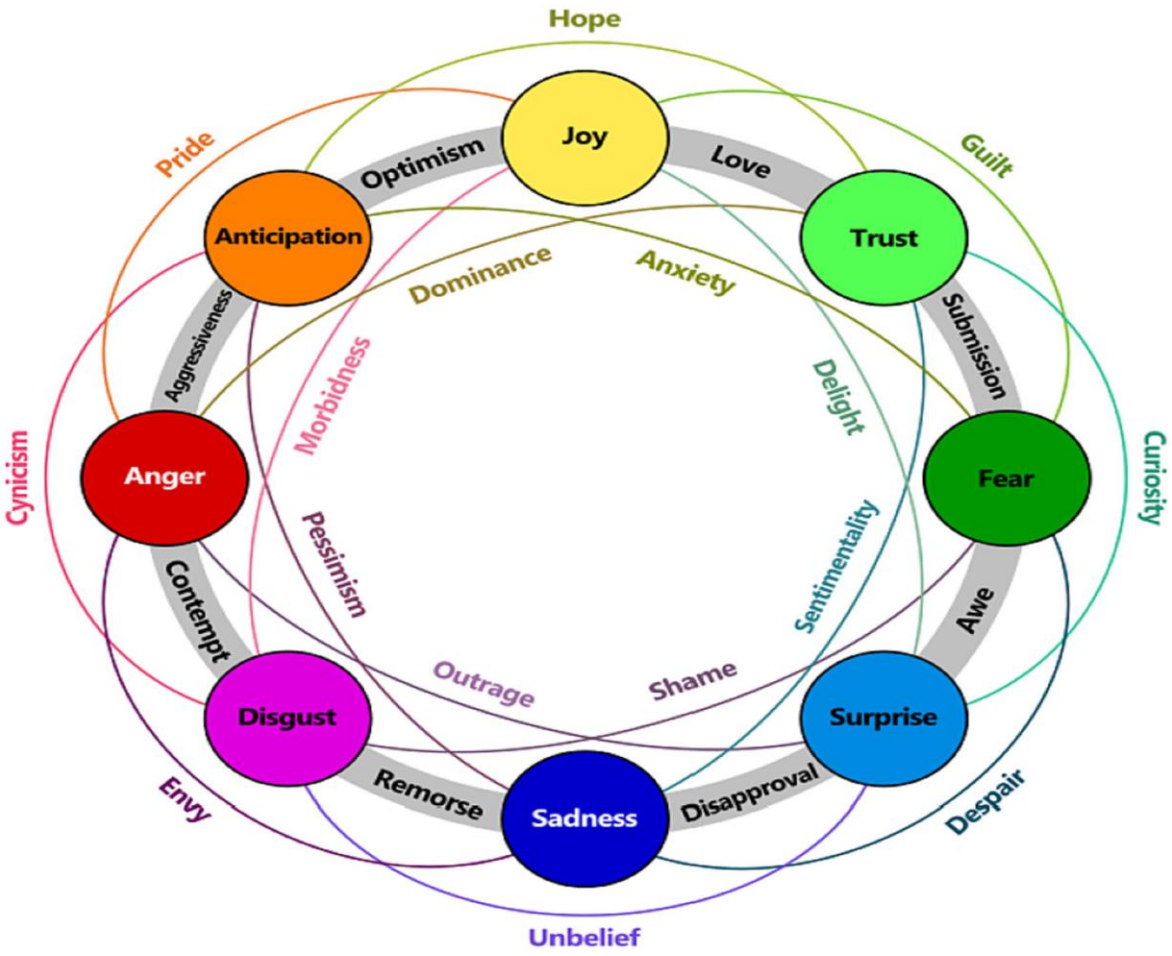
- Opinion spam is another problem inhibiting accurate sentiment analysis. Spam distorts product quality evaluation and precision of the polarity recognition of an opinion.

- Negations processing is used when negation before a word changes the polarity of a connected word. The most used negation processing methods are the switch and the shift negation.

- Emotion detection seeks to identify if a text expresses any type of emotion or not. Also, a problem of identification of the polarity of detected emotion is often necessary.

Extracting context from text is one of the most remarkable acquisitions obtained with natural language processing (NLP). A few years ago, context extraction was supposed to detect the polarity of sentiment from text, then the world took a step forward to detect sentiment in the form of emotions. These two concepts are very different. Sentiment can be positive, negative, or neutral, while emotions are more refined categories between positive and negative. Positive sentiment can be attributed to a happy, joyful, excited, and even funny emotion. Similarly, anger, disgust and sad emotions cause the sentiment to be negative. Several years ago, many machine learning algorithms were used in experiments to be training emotion detection models, but also to use a lexical approach for emotion recognition based on lexicons as lists of emotional words typical for specific emotions. However, all these approaches are slowly becoming obsolete due to the new trends in the deep learning detection models, which can do a very accurate automatic

analysis of emotions from a text. The most well- known and successful models being CNNs and recurrent neural networks (RNN), particularly LSTM.



A valuable survey of approaches of emotion recognition was made in which three major directions in emotions recognition were defined: categorical/discrete, dimensional, and appraisals-based approaches.

- Basic emotion model: the categorical approach - claims there are a small number of basic emotions that are hard-wired in our brain and recognized across the world. Each affective state is classified into a single category.

- Dimensional feeling model: the dimensional approach - based on the fact that feelings can be described as pleasantness–unpleasantness, excitement–inhibition and tension–relaxation. An example is Plutchik's wheel of emotions illustrated in figure above. In Plutchik's wheel of emotions, primary, secondary, and tertiary dyads are presented. Each dyad describes the distance between two emotions (hence the term dyads) as follows: the primary dyad

combines emotions next to each other, Joy – Trust; the secondary dyad combines two emotions when another emotion between them is skipped, Joy – Fear with skipped Trust, the tertiary combines two emotions when two other emotions between them are skipped, Joy – Surprise with skipped Trust and Fear.

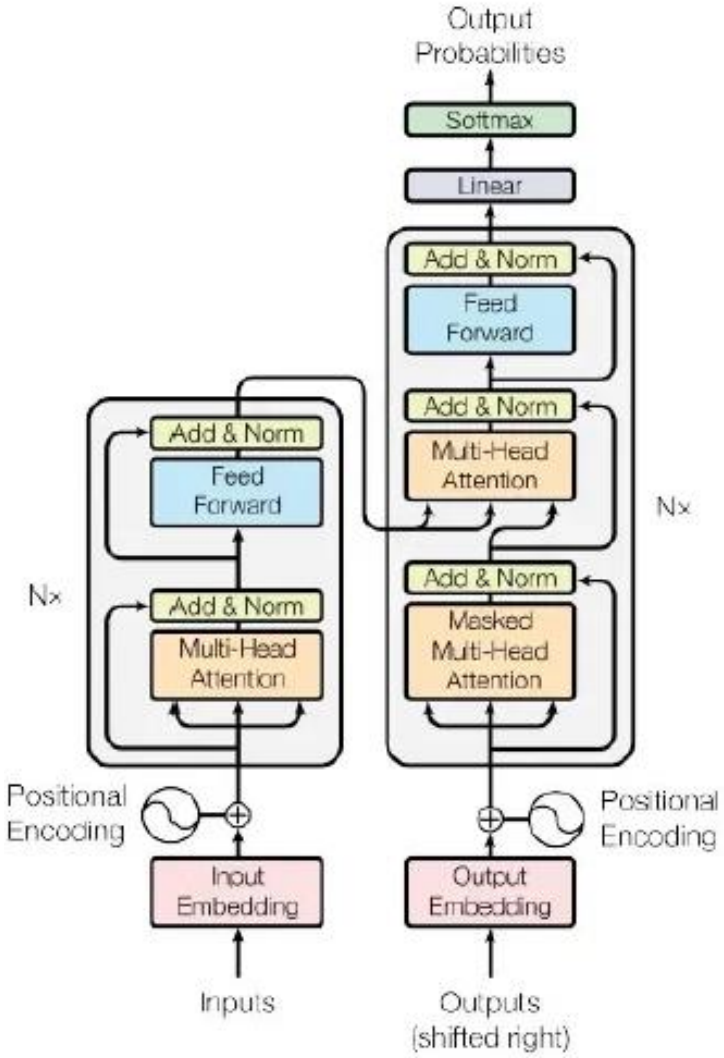
- Componential appraisal models: proposes that emotions are extracted from our appraisals (i.e., our evaluations, interpretations, and explanations) of events. These appraisals lead to different specific reactions of different people. It defines emotions as a balanced reaction to events, agents, and objects, and considers balanced reactions to differentiate between emotions and non-emotions. This approach is very suitable for affect sensing from the text.

Introduction to ChatGPT

ChatGPT is a language model developed by OpenAI, a leading artificial intelligence research organization. It is based on the transformer architecture, which has revolutionized the field of natural language processing. This model has been trained on a massive amount of data, allowing it to generate text and respond to various prompts with human-like precision and accuracy.

ChatGPT is based on Transformer architecture. It is a neural network architecture for processing sequential data, such as text. It was introduced in the 2017 paper "Attention is All You Need". The Transformer architecture is based on self-attention mechanisms, which allow the model to weigh the importance of different parts of the input sequence when making predictions.

The following is a detailed explanation of the Transformer architecture:



1. Input Sequence (Inputs): The input sequence is a sequence of tokens (e.g., words or sub-words) that represent the text input.

2. Input Embedding: The first step in transformation is to convert the input sequence into a matrix of vectors , where each vector represents a token in the sequence. This process is called input embedding. The input embedding layer maps each token to a high-dimensional vector that captures the semantic meaning of the token.

3. Self-Attention Mechanism: The self-attention mechanism allows the model to compute relationships between different parts of the input sequence . It consists of three steps: query, key, and value computations , and attention computation. In the query, key, and value computations, the input vectors are transformed into three different representations using linear transformations. In the attention computation step, the model computes a weighted sum of the values, where the weights are based on the similarity between the query and key representations. The weighted sum represents the output of the self-attention mechanism for each position in the sequence.

4. Multi-Head Self-Attention: The Transformer architecture uses multi-head self-attention, which allows the model to focus on different parts of the input sequence and compute relationships between them in parallel . In each head, the query, key, and value computations are performed with different linear transformations, and the outputs are concatenated and transformed into a new representation.

5. Feedforward Network: The output of the multi-head self-attention mechanism is fed into a feedforward network, which consists of a series of fully connected layers and activation functions. The feedforward network transforms the representation into the final output.

6. Layer Normalization (Add & Norm Layer): The activations in each layer of the Transformer architecture are normalized using layer normalization, which helps stabilize the training process and prevent the model from overfitting. A residual connection followed by layer normalization, which helps to stabilize the training process and make the model easier to train.

7. Positional Encoding: To capture the order of the tokens in the input sequence, a positional encoding is added to the input embedding. The positional encoding is a vector that represents the position of each token in the sequence.

8. Stacking Layers: The Transformer architecture can be stacked to form a deep neural network by repeating the multi-head self-attention mechanism and feedforward network multiple times.

9. Output: The final output from the Transformer, which is a vector representation of the input sequence.

During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. The term " in-context learning " to describe the inner loop of this process, which occurs within the forward-pass upon each sequence. The sequences in this diagram are not intended to be representative of the data a model would see during pre-training but are intended to show that there are sometimes repeated sub-tasks embedded within a single sequence.

Unlike traditional NLP models that rely on hand-crafted rules and manually labeled data, ChatGPT uses a neural network architecture and unsupervised learning to generate responses. This means that it can learn to generate responses without needing to be explicitly told what the correct response is, which makes it a powerful tool for handling a wide range of conversational tasks.

The model is trained using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. Initially the model is trained using supervised fine-tuning: human AI trainers provided conversations in which they played both sides — the user and an AI assistant. Trainers then access to model-written suggestions to help them compose their responses. It then mixed this new dialog dataset with the InstructGPT dataset, which transformed into a dialog format.

Key features and capabilities:

- Generating text and responses based on prompts
- Chatting and conversational AI
- Interactive storytelling.
- Content generation.
- Customer service and support

Innovative ChatGPT Tips and Tricks

ChatGPT can be customized and optimized to suit your needs. Here are a few tips and tricks to help you get the most out of ChatGPT.

1. Using creative prompts: The way your prompt ChatGPT can significantly impact the quality and creativity of its responses. Try using unconventional or creative prompts to see what kind of responses you can get.

2. Building chatbots: ChatGPT can be used to build chatbots for customer service, sales, and other applications. You can also try building chatbots with personality and character to enhance the user experience.

3. Generating content: ChatGPT can be used to generate various types of content, such as text, summaries, and even poems. Try using ChatGPT to generate content in different styles and formats.

4. Using ChatGPT for language translation: ChatGPT can be used for language translation and fine-tuning it on specific language data can help to improve its accuracy.

5. Using ChatGPT in combination with other models: ChatGPT can be combined with other models, such as GPT-3, to create even more advanced and sophisticated AI applications.

6. Experimenting with different APIs: OpenAI offers a range of APIs for working with ChatGPT, each with its own set of capabilities and limitations. Try experimenting with different APIs to see which one works best for your needs.

7. Personalizing the model: ChatGPT can be personalized to adapt to specific use cases and domains by fine-tuning it on relevant data. ChatGPT is a highly flexible model and fine-tuning it on specific datasets can help to improve its performance and accuracy.

8. Customizing responses with control codes: You can use control codes to customize the responses generated by ChatGPT. For example, you can use control codes to change the tone or style of the responses.

Limitations

1. ChatGPT sometimes writes plausible sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during RL training, there's currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.

2. ChatGPT is sensitive to tweaks to the input phrasing or attempting the same prompt multiple times. For example, given one phrasing of a question, the model can claim to not know the answer, but given a slight rephrase, can answer correctly.

3. The model is often excessively verbose and overuses certain phrases, such as restating that it's a language model trained by OpenAI. These issues arise from biases in the training data (trainers prefer longer answers that look more comprehensive) and well-known over-optimization issues.

4. Ideally, the model would ask clarifying questions when the user provided an ambiguous query. Instead, our current models usually guess what the user intended.

5. While OpenAI team made great efforts to make the model refuse inappropriate requests, it will sometimes respond to harmful instructions or exhibit biased behavior. OpenAI are using the Moderation API to warn or block certain types of unsafe content, but they expect it to have some false negatives and positives for now. They are eager to collect user feedback to aid ongoing work to improve this system.

Introduction to Google Bard

Google Bard is an impressive leap in language technology, building on the powerful (Bidirectional Encoder Representations from Transformers) BERT model. BERT is known for understanding context in language, and Google Bard takes it a step further. It's like a supercharged version that can tackle tricky tasks. Imagine asking your voice assistant a question using your natural way of speaking. Google Bard can handle that. It's good at understanding the words you say and the little things like how you tell them. This capability makes voice searches much more accurate and helpful.

Have you ever typed out a particular question? That's a long-tail query. Google Bard shines here too. It can read through the long sentences and figure out precisely what you're asking, giving you spot-on answers. In short, Google Bard is like the super-smart language wizard of AI. It's making conversations with computers feel smoother and more like chatting with a human. This feature is the kind of technology shaping how people talk to their devices and get the information they need.

Key Components of Google Bard AI

Google Bard AI's architecture comprises several vital components that synergistically contribute to its impressive performance in natural language understanding. These components allow the model to shine in understanding human language's context, nuances, and complexities.

1. Bidirectional Processing: Central to Google Bard's prowess is its utilization of bidirectional processing. Unlike traditional models that process language in one direction, Google Bard employs both forward and backward passes, allowing it to more effectively capture the relationships between words. This feature means it comprehends words based on what came before and what comes after, leading to a richer understanding of context. For example, it can differentiate between "let him go" and "go let him."

2. Attention Mechanisms: The attention mechanism is another integral aspect of Google Bard's architecture. This mechanism allows the model to assign varying levels of importance to different parts of a sentence while considering the entire context. By doing so, Google Bard can focus on crucial words and phrases, accounting for their significance in the overall meaning.

3. Contextual Embeddings: Google Bard employs contextual embeddings, representations of words that incorporate contextual information. These embeddings capture the word's meaning based on the surrounding terms, ensuring that words with multiple meanings are correctly interpreted within the given context. This way, Google Bard can better grasp the overall purpose of a sentence, leading to more accurate interpretations.

Benefits of Google BARD AI

Google Bard AI brings forth remarkable benefits from its advanced language understanding capabilities. Here are a few of its standout features:

1. Enhanced Contextual Understanding: With bidirectional processing and attention mechanisms, Google Bard understands the intricacies and context of language. That means it can accurately grasp what you mean, even in complex queries or when you use ambiguous phrases.

2. Accurate Voice Searches: Google Bard is fantastic at understanding natural speech patterns. It excels in delivering accurate responses during voice searches. You'll love how it grasps the subtleties of spoken language, making voice assistants more reliable and enhancing your overall user experience.

3. Precision in Long-tail Queries: When users input detailed or lengthy queries, Google Bard's contextual embeddings shine. It captures the exact intent behind these queries, providing precise answers matching the user's needs.

4. Improved Search Relevance: Google Bard enhances search engines' ability to deliver relevant results. It understands context, eliminating confusion caused by polysemy (multiple-word meanings) and ensuring users receive information tailored to their intended context.

5. Personalized Content Generation: The model's understanding of context enables it to generate contextually relevant content. You can harness it for customized recommendations, content summaries, and creative writing.

The underlying technology behind the AI chat service associated with Google Bard is Google's Language Model for Dialogue Applications (LaMDA). LaMDA was revealed two years prior, showcasing Google's progressive approach to language technology development. Google Bard leverages the advancements of LaMDA to create a new realm of AI-driven conversation, interaction, and comprehension. It opens up exciting possibilities in natural language understanding and engagement.

What is Google Bard used for?

Thanks to its advanced language understanding capabilities, Google Bard finds applications across a spectrum of domains.

1. Search Engines Enhancement: Google Bard enhances search engines' ability to comprehend complex queries, delivering more accurate and contextually relevant results. This feature is especially crucial for ambiguous search terms where context plays a pivotal role.

2. Voice Assistants: In voice searches and interactions, Google Bard accurately interprets spoken language nuances. It enables voice assistants to provide precise and relevant responses, improving user experience.

3. Natural Language Processing (NLP) Tasks: Google Bard's bidirectional processing and contextual understanding empower it to excel in various NLP tasks like sentiment analysis, text classification, and named entity recognition, aiding in automating language-related tasks.

4. Content Generation: Leveraging its contextual embeddings, Google Bard can generate contextually relevant content, personalized recommendations, summaries, or creative writing, streamlining content creation processes.

5. Conversational AI: Google Bard's robust comprehension abilities are instrumental in developing advanced chatbots and conversational agents, enabling more natural and intuitive interactions between humans and machines.

Is Google Bard better than ChatGPT?

When comparing Google Bard and ChatGPT, it's important to consider their unique strengths and focuses. Google Bard is designed to excel in understanding natural language. With bidirectional processing and contextual embeddings, it can effectively interpret complex queries, enhance voice searches, and improve search relevance. It's particularly valuable for applications such as voice assistants, search engines, and content generation.

ChatGPT, on the other hand, is renowned for generating human-like text and engaging in versatile conversations across a wide range of topics. It's designed to be a conversational partner and creative content generator, making it suitable for content creation, brainstorming, and interactive dialogue tasks.

Ultimately, the choice between Google Bard and ChatGPT depends on the specific application. If language understanding and query interpretation are the primary needs, Google Bard is a strong contender. For engaging and creative conversations, ChatGPT remains a preferred choice. Both technologies represent significant advancements in AI, catering to different aspects of human-computer interaction.

How does Google Bard work?

Users can use Google Bard by signing up for or logging into a Google account. Once logged in, they can access the conversational AI service directly. Users input their queries, ranging from voice searches to text-based questions. Google Bard then employs its language comprehension capabilities to provide accurate, contextually relevant responses, creating a seamless and intuitive interaction between users and AI-driven language understanding.

Who has access to Google Bard?

Google Bard is available to everyone with a Google account. That means users from all over the globe, representing different cultures and backgrounds, can now access AI-driven language understanding capabilities. In this way, Google Bard transcends the traditional barriers of language and comprehension, bridging the gap between humans and machines.

Does Google Bard include images in its answers?

Yes, Google Bard can provide images in its answers. When users input image-related queries, Google looks for relevant content within its vast database of images and delivers the results accordingly. For example, when you ask, "What does a golden retriever look like?" it returns pictures of the breed as an answer.

Challenges and Limitations of Google BARD

Google Bard is a remarkable step forward in natural language understanding and human-computer interaction. Still, it is not perfect. Some challenges and limitations come with the technology.

1. Limited Domain Understanding: Google Bard is tailored for general conversation, drawing from a wide range of topics and contexts. However, its domain understanding capabilities are limited to the broader scope. It's still unable to pick up on specialized, technical phrases and terms that require more in-depth knowledge of a specific domain.

2. Learning Curve: For novice users, a learning curve is involved with using Google Bard. The AI understands conversational language but falls short on more intricate queries. As such, users need to understand the model's capabilities and limitations, as well as how to correctly phrase their questions.

3. Limited Accuracy: Although Google Bard is good at understanding natural language patterns, it isn't perfect at identifying every nuance of a query. It tends to be less accurate in complex queries with many variables or when words have multiple meanings within the same context.

Future Implications and Evolution

Google Bard is set to revolutionize human-computer interaction through its advanced language understanding capabilities. In the years to come, it will continue to evolve and further refine its comprehension and conversational AI performance. Here are a few potential developments:

1. Enhanced Domain Understanding: To accommodate specialized conversations, Google Bard will likely advance its understanding of specific domains. This feature would enable it to grasp technical phrases and terms better, expanding its applicability for more in-depth conversations.

2. Improved Accuracy: As Google Bard continues to mature, its accuracy is also set to increase. With improved language models, the AI can learn to interpret complex queries more precisely, leading to more satisfactory outcomes for users.

3. Expanded Usage: As Google Bard evolves, more industries and applications will find use for the AI's language understanding capabilities. It can be used for natural language processing tasks, content generation, automated customer service processes, and interactive dialogue systems.

Explainable AI

A distinguishing feature of today's AI is not limited to observable methods, and, when it reaches a certain level of complexity, it poses interpretability challenges. In other words: AI models tend to have a high performance, much higher than traditional algorithms; but in each specific case it can be extremely complex to explain why the model has produced a given result. Although there are applications of AI where it is not as important to be able to understand or explain why the algorithm has returned a particular value, in many cases it is essential and is a regulatory requirement.

All this has led to the development of the Explainable Artificial Intelligence (XAI) discipline, which is the field of study that aims to make AI systems understandable to humans, as opposed to the notion of “black box”, which refers to algorithms in which only the results are observable and the operation of the model is unknown, or the basis for the results cannot be explained.

It can be concluded that an algorithm falls within the XAI discipline if it follows three principles: transparency, interpretability and explainability. Transparency occurs if the processes that calculate the parameters of the models and produce the results can be described and justified. Interpretability describes the ability to understand the model and present how it makes decisions in a human-understandable way. Explainability refers to the ability to decipher why a particular observation has received a particular value. In practice, these three terms are closely linked and are often used interchangeably, in the absence of a consensus on their precise definitions.

These principles are achieved through basically two strategies: either develop algorithms that are interpretable and explainable by their nature (including linear regressions, logistic or multinomial models, and certain types of deep neural networks, among others), or use interpretability techniques as tools to achieve compliance with these principles

XAI deals both with the techniques to try to explain the behavior of certain opaque models (”black box”) and the design of inherently interpretable algorithms (”white box”). XAI is essential for AI development, and therefore for professionals working in this area, due to at least three factors:

- It contributes to building confidence in making decisions that are based on AI models; without this confidence, model users might show resistance to adopting these models.
- It is a regulatory requirement in certain areas (e.g. data protection, consumer protection, equal opportunities in the employee recruitment process, regulation of models in the financial industry).
- It leads to improved and more robust AI models (e.g. by identifying and eliminating bias, understanding the relevant information to produce a certain result, or anticipating potential errors in observations not included in the model’s training sample). All of this helps to develop ethical algorithms and allows organizations to focus their efforts on identifying and ensuring the quality of the data that is relevant to the decision process.

Context and rationale for XAI

1. Digital transformation has enabled access to and exploitation of a vast amount of structured and unstructured data, driving the use of machine learning techniques and artificial intelligence across industries.

2. AI models provide greater predictive power, but they also present risks, such as the presence of undetected bias, lack of understanding of the model, or errors in its application arising from causes such as overfitting, all of which can lead to model distrust. This raises the question of whether it is possible to understand the results of AI algorithms well enough to make appropriate decisions.

3. Explainable Artificial Intelligence (XAI) is a set of processes and methods that enable users to understand and trust the results and products created by machine learning algorithms. This discipline is crucial for an organization to build trust when using AI models, helping to characterize model accuracy, fairness, transparency and understanding of results in AI-based decision making.

4. Academic and business interest in XAI has increased exponentially in recent years, due to this discipline's ability to address a number of industry concerns regarding the use of AI, such as regulatory requirements, lack of trust, potential misuse, reputational impact, social or human impacts, and other risks.

5. This has led regulators and supervisors in different jurisdictions to establish regulations and guidelines for the appropriate use of AI, including the interpretability aspects of models.

6. In Europe, the European Parliament's General Data Protection Regulation (GDPR) that came into force in 2018 includes a "right to an explanation" for citizens, requiring companies to be able to explain why an AI model yielded a certain result. This has critical implications for the design and interpretability analysis of AI models.

7. Moreover, in 2021 the European Parliament proposed the Artificial Intelligence Act (AI Act) to regulate the use of artificial intelligence in the European Union. This proposed Regulation sets out a regulatory framework for AI systems, including requirements for ethical development, transparency, security and accuracy, as well as a governance and oversight system. The AI Act classifies AI applications into levels of risk (unacceptable practices, high-risk systems, and low or limited risk systems), and lays down transparency and human oversight requirements for high-risk systems, which will be enforceable across the Union. This is likely to trigger initiatives to adapt to the Regulation, including comprehensive model documentation, interpretability techniques, monitoring dashboards and model alerts.

8. Likewise, in 2019 the European Commission formulated the Ethical Guidelines for Trustworthy Artificial Intelligence, which propose seven key requirements for AI systems to be considered trustworthy: (i) human agency and oversight, (ii) technical robustness and safety, (iii) privacy and data governance, (iv) transparency, (v) diversity, non-discrimination and fairness, (vi) social and environmental well-being, and (vii) accountability. The transparency requirement includes the need for AI models to be explainable. The Guidelines propose evaluation criteria to assess the extent to which an AI model meets these requirements.

9. In the United States, the White House proposed an AI Bill of Rights in 2022, pushed by President Joe Biden. This bill sets out five principles or citizen rights regarding AI, including safe and effective systems, protection against discrimination by algorithms, data privacy, notification and explanation, and evaluation and correction by a human in the event of AI failure (fallback). These principles include the explainability of AI models, which requires plain language documentation in addition to technically valid, meaningful and useful explanations, and demonstrably clear, timely, understandable and accessible notices of use.

Definition

The XAI discipline is relatively new, and therefore there is not yet a settled doctrine that standardizes its terminology. Despite some notable efforts to define terms, the approach to XAI is either diverse (depending on the academic source consulted) or intuitive (more frequently in industry).

In any case, for most uses in practice it may be sufficient to define XAI as follows:
Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms. Explainable AI is used to describe an AI model, its expected impact and potential biases. It helps characterize model accuracy, fairness, transparency and outcomes in AI-powered decision making. Explainable AI is crucial for an organization in building trust and confidence when putting AI models into production. AI explainability also helps an organization adopt a responsible approach to AI development.

Relevance of XAI

One aspect on which there is consensus among academics and industry professionals is the growing relevance of XAI as a complementary discipline to AI. Scientific publication analysis tools identify more than 77,000 articles on XAI between 2014 and 2022, and this trend is exponentially increasing, with more than 20,000 articles in 2022 alone.

Beyond academic interest, the attention XAI receives is explained by its ability to provide solutions to industry concerns around the use of AI including:

- Regulatory requirements: the obligation to comply with emerging regulations on the use of AI.

- Lack of confidence: the need to build confidence in the AI model and the results it delivers among users, validators and auditors, and ultimately the general public.

- Potential misuse: the desirability of avoiding misuse of the models due to lack of understanding of how they work, which can lead to costs and even penalties.

- Reputational impact: the prevention of reputational impacts for organizations due to model bias, discriminatory decisions, erroneous predictions by the model or inappropriate use.

- Social or human impacts: the prevention of harmful social or human impacts in critical uses such as AI for the diagnosis of medical diseases, judicial sentences, biometric identification, polygraphs, etc.

- Other: mitigation of other risks arising from lack of understanding about the model, such as cybersecurity, data protection, fraud, model risk.

Artificial Intelligence Act (European Parliament)

The draft Artificial Intelligence Regulation or Artificial Intelligence Act (AI Act), published in 2021, is a proposal for the use of artificial intelligence in the European Union that aims to ensure a high level of trust in AI and its applications, while laying the groundwork for innovation. The Regulation establishes a regulatory framework for AI systems in the EU, and includes requirements for ethical development, transparency, security and accuracy. It also establishes a governance and oversight system for AI systems, as well as data protection and data governance rules.

As it is a Regulation, when approved, it will be directly applicable in the Union's 27 countries without the need to be transposed into each country's legal system. One of its key features is that it sorts AI applications into risk levels:

- Prohibited practices is the highest risk category and systems falling under this category are totally forbidden. They include:
  - Real-time remote biometric systems that can be used for any type of surveillance, although exceptions apply for crime prevention and criminal investigations in law enforcement and homeland security contexts.
  - Social scoring algorithms that can be used to evaluate individuals based on predicted personal or personality characteristics leading to detrimental or unfavourable treatment of an individual.
  - Subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm.

- High-risk AI systems is likely to constitute the majority of AI systems. These include:
  - Biometric identification and categorization of natural persons
  - Management and operation of critical infrastructure [e.g. traffic].
  - Education and vocational training
  - Employment, workers management and access to self-employment
  - Access to and enjoyment of essential private services and public services and benefits, including creditworthiness assessment, credit rating or prioritization of access to such services (Note: this aspect applies to AI systems used in the financial services sector in particular).
  - Law enforcement
  - Migration, asylum and border control management
  - Administration of justice and democratic processes

- Low-risk (or limited-risk) IA systems, covering systems that do not use personal data or make predictions that could affect individuals directly or indirectly, such as industrial predictive maintenance applications.

Regarding the interpretability of AI models classified as high risk, the AI Act establishes in its Articles 13 and 14:

Art. 13. Transparency and provision of information to users.

1. High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately.

2. High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users.

Art. 14. Human oversight

1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.

2. The measures referred to shall enable the individuals to whom human oversight is assigned to do the following, as appropriate to the circumstances:

a. fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible.

b. remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (‘automation bias’).

c. be able to correctly interpret the high-risk AI system’s output.

d. be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system.

e. be able to intervene on the operation of the high-risk AI system or interrupt the system.

As can be seen, the AI Act imposes restrictive conditions on the interpretability of high-risk AI models, which will soon become mandatory throughout the Union. This is expected to trigger a significant number of initiatives to adapt to the Regulation, including more exhaustive documentation of models and their uses, the implementation of interpretability techniques, the development of model monitoring and alert dashboards, and a review of the full model development, validation, implementation and use procedure.

Ethical Guidelines for Trustworthy Artificial Intelligence (European Commission)
In April 2019, the European Commission's High Level Expert Group on AI presented the Ethical guidelines for trustworthy AI29, following a consultation process with more than 500 industry responses.
The Guidelines propose seven key requirements that AI systems must meet to be considered trustworthy, which in summary are:

(i) human agency and oversight

(ii) technical robustness and safety

(iii) privacy and data governance

(iv) transparency

(v) diversity, non-discrimination and fairness

(vi) social and environmental well-being

(vii) accountability

Bias in Artificial Intelligence

A machine can't have bias, right? After all, it doesn't have experiences or memories from which to form said bias. Unfortunately, that's not quite the case: machines can only learn from the data they have and if this data is biased, incomplete, or of poor quality, the output of the machine will reflect the same problems.

The following are the most common examples of artificial intelligence bias:

- Algorithm bias: if the algorithm itself that determines the calculations of the machine are incorrect or faulty, the results will be as well.

- Sample bias: if the dataset you select doesn't accurately represent the situation, your results will reflect this error.
  Example: you're collecting salary information, but only record those of male employees.

- Prejudice bias: similarly to sample bias, prejudice bias uses data that is influenced by societal biases and therefore incorporates this prejudice into what should be opinion-free data.
  Example: you're evaluating the gender distribution in certain occupations, but only count female teachers and male doctors, creating an inaccurate skew in your data.

- Measurement bias: measurement bias occurs when data is incorrectly gathered, specifically on how it was measured or valued.
  Example: if employees are surveyed about their feelings about their employer and promised a reward if enough employees answer, those who are motivated simply by the reward may not give thorough or accurate responses.

- Exclusion bias: you can't pick and choose the data you use in your analysis and if you (intentionally or by mistake) exclude data points, your results will be inaccurate.
  Example: if you think the middle-of-the-road answers to a survey aren't consequential and remove them, you'll end up with data skewed to both ends of the spectrum and an inaccurate representation of how the respondents actually feel.

- Selection bias: while it can be quite challenging to get a big enough sample or one that's representative of the entire population, choosing only certain groups can make your data completely useless.
  Example: you want to evaluate the universities that high school graduates choose to attend, but ignore those who choose to immediately enter the workforce or attend community college, therefore painting an inaccurate picture of your graduates' choices.

There are quite a few more ways that bias can appear in artificial intelligence, but the aforementioned ones are the most common. Here's what you need to remember: artificial intelligence learns from the data that it's fed and if that data is problematic or inaccurate, the outputs of artificial intelligence will be as well. Here's what you can do to prevent bias:

1. Lots of situations involving bias stem from small or limited datasets; do everything you can to collect as much data as possible from as many sources as you can, diversifying your dataset.

2. As you begin to feed your computer with data, run tests during the early stages of testing to check for biases and correct them.

3. Explore online fairness and bias tests to make sure you caught everything.

4. Run your results by other experts to get other opinions and continuously check the quality of your data as time passes.

Ethics in Artificial Intelligence

You've definitely heard someone tell you that AI will take your job one day. While the vast majority of jobs are safe (and those that AI can take over will morph into a different role), there are serious ethical considerations to keep in mind when discussing artificial intelligence.

One thing is clear: the power of artificial intelligence is massive and we've only just begun to uncover what it can do. But the following considerations are absolutely crucial when it comes to maintaining proper ethics in the future of artificial intelligence:

- Privacy: we're feeding machines tons of data about people to help it react in a more human-like way, right? How do we ensure that the data we're giving to the machine is both secure and private? Prioritizing data privacy throughout the entire artificial intelligence lifecycle is one of the world's main concerns.

- Human dependence: yes, artificial intelligence is capable of automating some tasks that humans were previously handling and it can also handle much more data than people can. But it's absolutely essential that AI isn't left to make decisions on its own, as it will never replace human responsibility and accountability.

- Sustainability: advances in artificial intelligence and technology are supported, but as long as they don't come at the expense of the environment and overall sustainability.

- Accessibility: new developments should be accessible worldwide, not just in highly developed countries with easy access to technology.

To ensure that ethics in artificial intelligence is prioritized, many countries and global organizations have come together to come up with policies and regulations, such as the GDPR in the European Union. But achieving truly ethical technological advances in artificial intelligence will come from a commitment from every individual, company, and country across the world. The power of artificial intelligence is truly unmatched–but it's on us to properly use it for good. And skilled artificial intelligence professionals are sorely needed across the tech industry, so if you're interested in entering this up-and-coming field, look no further: there's lots of room for advancement in artificial intelligence.

Societal Impact of AI

The rapid advancement of Artificial intelligence (AI) and the increasing prevalence of AI in business and society will have a sweeping impact on the future of jobs and the future of society. According to a Forrester report, generative AI is projected to displace approximately 2.4 million jobs in the US, along with other forms of automation replacing additional positions. Studies predict that AI will push many out of work and create a "useless class". New economic models, social systems, and education re-engineering are necessary in the AI era. Research on the societal impact of AI is critical and urgently needed.

While AI provides much potential, its application is still lingering with issues. For example, the operation of AI models generally needs a large training data set, which may necessitate the incorporation of external data. These external data acquisitions will require appropriate procedures and systems to effectively handle the integration with existing data sets. Errors will occur without proper governance. Also, the decision processes are complex and invisible. The transparency issues make it difficult to convince people to trust recommendations provided solely by machines. Illegal and unethical consequences may happen since AI algorithms could identify relationships that are not causal, resulting in biases against certain groups of people. Further, the integration of AI into the workforce may put some traditional jobs at risk.

First, we are studying how the introduction of AI challenges existing legal and ethical standards. Legal standards are mandatory and must be observed. Ethical ones indicate what is accepted by society. People observe rules adopted by the majority as a social obligation. Thus, ethics can be viewed as internal systems of control, whereas law refers to external mechanisms of enforcement.

The rapid development of AI is raising a series of legal and ethical concerns. For example, the introduction and application of AI may result in people losing their jobs, which may cause social instability. In healthcare, the use of autonomous robotic devices has aroused significant concerns

about ethics and trust. People are worried that AI can harm human physical and mental integrity and reduce human autonomy. The existence of such problems is partially because of the lack of legal and ethical frameworks related to AI, and previous research studies have not dealt with the issue comprehensively. Value-Focused Thinking is used which is a systematic qualitative method, to conduct the study. Ideas and opinions on AI development are collected from both IS/IT professionals and legal professionals. The preliminary results show that experts believe that "maximize ethical AI development" and "maximize AI governance" are fundamental. To achieve these two fundamental objectives, different levels of means objectives, such as "maximize clarity of AI liability", "maximize communication", and "maximize social stability", are needed.

A key ethical AI issue is related to bias and discrimination. Bias and discrimination issues derive from both technical issues and human-related issues. Technically, the over-and under-representativeness of the data used in AI models may lead to minority bias as certain groups are not fully considered. Data cleansing is a procedure of spotting and correcting inaccuracies in data to increase its quality. Prioritizing the enhancement of data quality can tap into the full potential of AI. Bias and discrimination results can be caused by unprofessional behaviors during the training processes. For example, inconsistencies in data labeling and unethical actions may occur while developers manually divide possible values of a target variable into exclusive categories. The algorithms used by AI models can also be problematic as it is hard for developers and users to find all the errors and biases in AI algorithms, which are usually in the "black box".

Humans are another factor in why bias and discrimination problems occur. One of the most important groups is the developers. Developers may have insufficient knowledge of social science and pay less attention to fairness issues. Another unsettled problem of this group is that it lacks diversity. A large proportion of AI developers are males, and the male-thinking model will place females in a disadvantageous position. A study was conducted aiming to detect how bias and discrimination issues affect the "collaboration" between AI users and AI applications. The results of this research will enable the developer to formulate better approaches to enhance users' trust in AI systems and enable human-AI collaboration to generate greater efficiencies and create more positive societal impacts.

Another study is the governance of AI. Researchers, policymakers, and all stakeholders need to pay extra emphasis on the governance of AI to maximize its positive impact on society. Many authorities, such as the United States, the European Commission, Singapore, and Hong Kong, have proposed frameworks for AI governance. One potential solution to tackle AI challenges, such as those related to ethical and legal issues, is to develop an integrated AI framework based on the meta-synthesis of existing AI frameworks from various regions worldwide. Such a unified framework will be structured by the core components which are identified from the existing frameworks. The unified governance framework that will be developed in this research can be specialized to different geographical regions, countries, and businesses based on the specificities, needs, and requirements of different scenarios and environments.

Since AI is bringing and will continue to bring significant societal impacts on the entire society and impact on the future of humanity, scientists, professionals, and even ordinary people are calling for regulations and policies to govern AI development and use. Researchers agree that determining the acceptable uses of AI is urgent. Regulations are needed to deal with problems including machine bias, legal decision-making, and legal responsibility. Machine-learning algorithms should be more explainable by enhancing the transparency of the input data, the testing of algorithms, and the decision model. The understanding of fairness, accountability, transparency, and explainability in AI systems is vital in boosting users' confidence in these systems. Users can gain deeper insights into AI's 'black box' by understanding how AI reaches its decisions. For instance, the systems can offer additional details, such as the size and limitations of the training set, to enhance comprehensibility.

Future research can explore who should be held accountable for AI decisions in different scenarios and the establishment of transparency and privacy agreements for disclosing necessary information. Many countries and geographical regions are proposing relevant documents to increase the trustworthiness of AI. For example, China released the "New Generation Artificial Intelligence Development Plan" in 2017. This strategy reveals China's vision to become a leader in AI by the end of the next decade and highlights ethical norms and standards for AI. On July 10, 2023, China's top internet watchdog, the Cyberspace Administration of China (CAC), in consultation with six other regulators, issued China's first generative AI regulation to legislate a largely unregulated AI space.

Engineers and scientists have signed a global pledge against the development and usage of autonomous weapons systems using AI that can identify, target, and kill a person without human authorization. They also advocate stopping generative AI development, as such technologies may threaten humanity. Many companies do not want too many regulations on AI as these restrictions may impede technology development and deprive them of the opportunities to gain competitive advantages. Ethical AI development and the need for AI governance are critical topics at this expeditiously changing and transformative time.

Interdisciplinary collaboration

Applying AI to different domains, such as accounting, finance, and healthcare, needs support from professionals in these fields. Insufficient collaboration among disciplines may lead to errors and inefficiencies. For example, AI developers' lack of knowledge and sensitivity to social science issues is one of the cited reasons why biases and discrimination occur. Domain-specific knowledge is essential for ChatGPT to provide convincing responses when it is used in various subjects. There is a call for speeding up the integration of large-scale models such as ChatGPT, a large language model, into healthcare. Healthcare areas such as diagnosing diseases, spotting malignant tumors, and drug discoveries are tasks that AI can do well and may perform better

than healthcare professionals. AI-healthcare professionals' collaboration is important. Besides the smooth operation of the AI systems, professional opinions and proper interpretation of the feedback from the clinics are critical for the realization of AI-empowered Medicare. Studies show that combining AI with human evaluations can maximize diagnostic accuracy and provide more optimal treatment planning. Similarly, experts familiar with tasks in fields such as accounting and finance are also important to ensure that the AI systems function well in those domains. With AI penetrating almost every field and discipline, interdisciplinary collaboration and cooperation will maximize the positive impact of AI on society.

Education reform

Society needs to invest more resources in education on the awareness, understanding, and usage of AI. Northeastern University proposed an educational model consisting of three components – technology, data, and human interaction. Society, either schools and colleges or working institutions, can provide students and junior workers with training in these three aspects. Specifically, they need to be equipped with sufficient skills and knowledge, such as machine learning, natural language processing, and deep learning, to face an era with rapidly evolving technologies, and they need to know how to use content generated by AI properly. Studies show that low-skilled laborers will be most negatively affected by the introduction of AI, while medium and high-skilled laborers may benefit from working with AI. Higher education institutions play essential roles in developing "soft skills", such as creativity, problem-solving, collaboration, communication, and adaptability, that enable students to be prepared for the AI era. The combination of theoretical knowledge and practical experience is also essential for professionals to transform into a new generation of professionals who can work collaboratively and in partnership with AI. In an era with a rapidly and frequently changing environment, the adaptability quotient (AQ) will be critical and needs to be nurtured by the educational system. The AI era has arrived. One needs to continuously adapt and transform oneself to adapt to and excel in the new era. Resistance is futile.