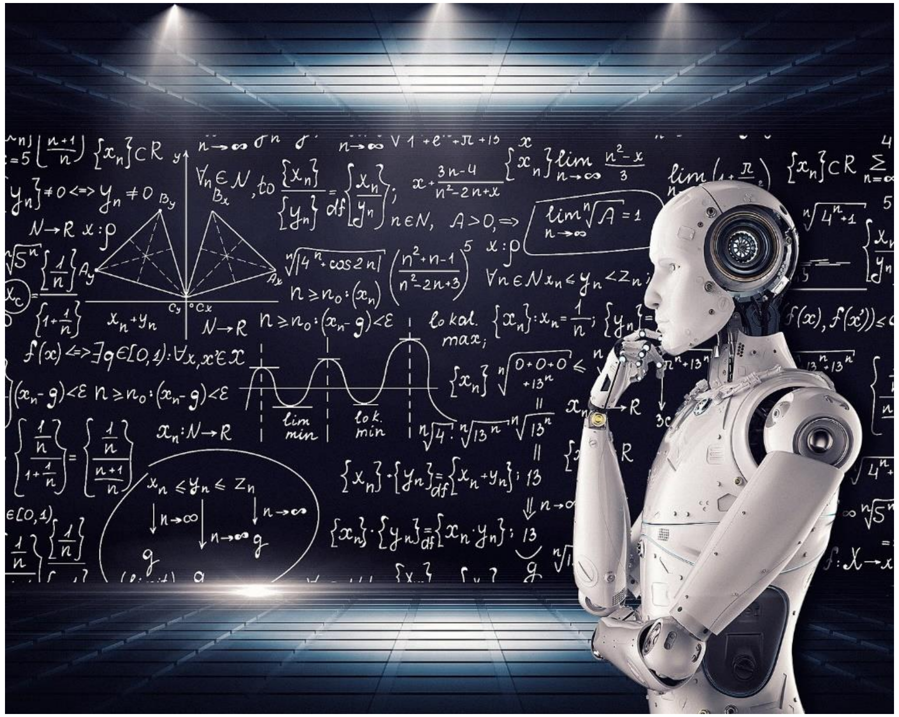
Generative and Conversational AI for B&IM

Peer-Olaf Siebers. University of Nottingham. UK. peer-olaf.siebers@nottingham.ac.uk



Picture credit: https://www.vpnsrus.com/

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Machine Intelligence: Fields

- Symbolic AI (around since mid 1950s)
  - Uses symbols (concepts) and logical reasoning to represent knowledge and solve problems
    - Examples: Expert systems; natural language processing; robotics
- Machine Learning (around since 2010s)
  - Allows computers to learn from data without being explicitly programmed
    - Examples: Image recognition; recommender systems
- Generative AI (around since 2020s)
  - Concerned with learning from existing data and then creates entirely new content
    - Examples: GPT-4 (general) DALL-E2 (text-2-image)
- Conversational AI (around since 2020s)
  - Allows computers to communicate/interact with humans in a natural language conversation
    - Examples: Alexa; ChatGPT

Generative AI Technology Stack

Application Layer

APIs and SDKs provide easy access to generative AI models for developers (e.g. OpenAI API; Gemini API)
Application Development Tools help developers to build the user interface (e.g. Flask)
End User Applications are the final products that users interact with (Chatbots (e.g. ChatGPT UI; Gemini UI); Content Creation (e.g. Midjourney; Synthesia); Virtual Assistants (e.g.Amazon Alexa; Google Assistant) Legal Apps (e.g. LexMachina); Financial Apps (e.g. AlphaSense) Code Generation Apps (e.g. GitHub Copilot))

Model Layer

Foundation Models are trained on massive amounts of heterogeneous data (e.g. GPT-4; Gemini)
Specific Models result from training foundation models on large datasets of specific content types (e.g. ChatGPT (conversation); DALL-E2 (text-to-image))
Specialised Models are tailored to specific domains, use cases or customers (ClaraHealth (healthcare domain); LegalSifter (contract review use case))
Model Hubs act as a central repository for accessing pre-trained models developed by various organisations or researchers (e.g. HuggingFace)

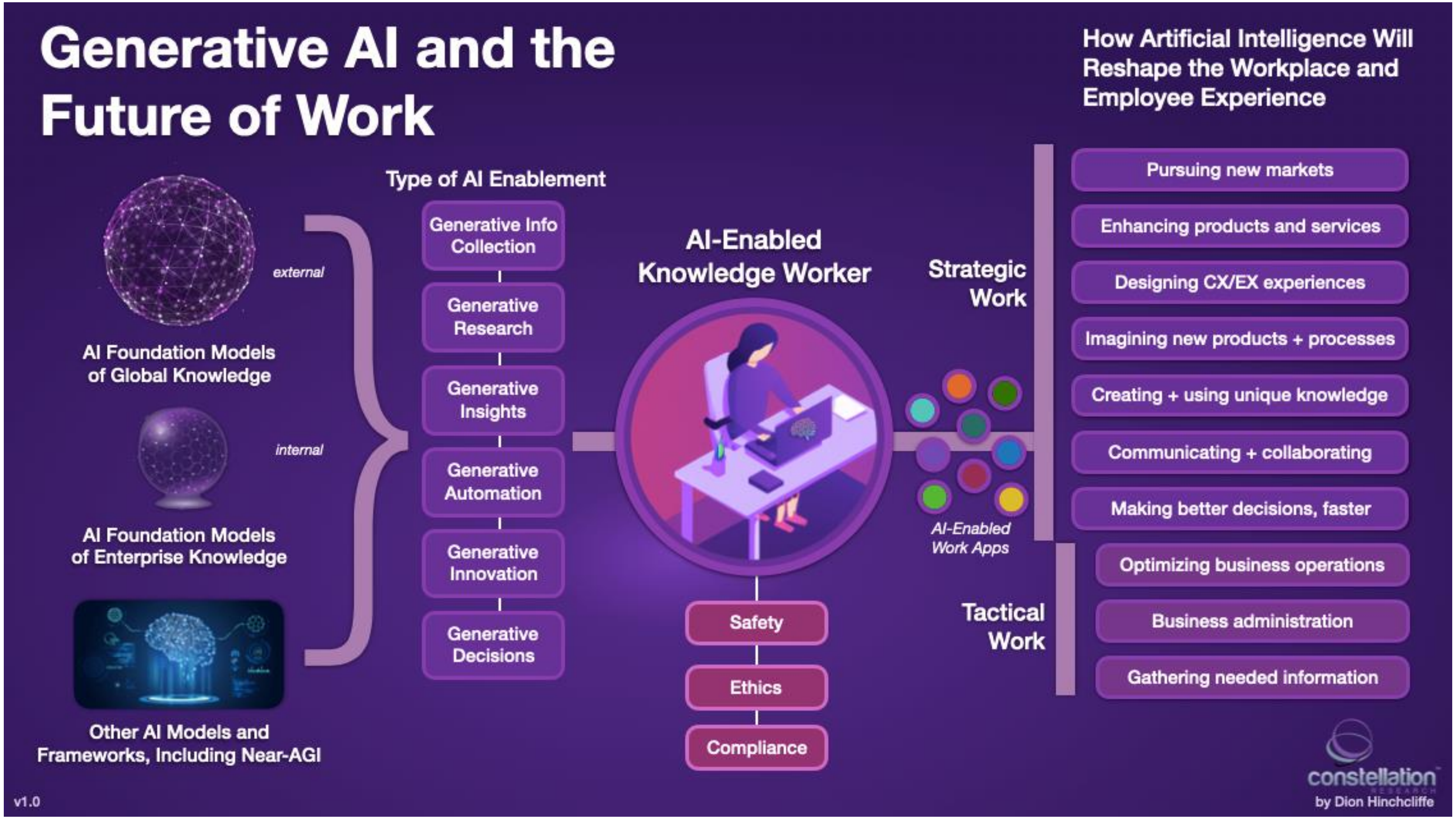
Infrastructure Layer

Hardware Platforms equipped with optimised chipsets (Graphics Processing Units (GPUs) from Nvidia or AMD, Tensor Processing Units (TPUs) from Google)
Cloud Platforms as remote delivery systems for computing resources (e.g. Google Cloud TPUs, Amazon Elastic Compute Cloud (EC2), Microsoft Azure)
Storage Platforms: High-performance storage systems from companies like NetApp, Dell EMC
Connectivity Platforms (wired and unwired) providing access to the cloud (e.g. BT, Virgin Media, Sky Broadband, TalkTalk)

Note: This is work in progress!

- Model types
  - Foundation Models: Trained on massive amounts of heterogeneous data (e.g. GPT-4)
  - Specific Models: Result from training foundation models on large datasets of specific content types (e.g. ChatGPT (conversation); DALL-E2 (text-to-image); BERT (natural language understanding); FaceNet: (facial recognition and verification)
  - Specialised Models: Tailored to specific domains, use cases or customers (ClaraHealth (healthcare domain); LegalSifter (contract review use case); JPMorgan Chase AI Legal Assistant (client-centric legal assistant))
  - Model Hubs: Act as a central repository for accessing pre-trained models developed by various organisations or researchers (e.g. HuggingFace)
- Model size
  - Large Foundation Models: Extensive, pre-trained models on diverse and heterogeneous data, serving as a base for various downstream tasks (e.g. GPT-4)
  - Small Foundation Models: Compact versions of large foundation models with fewer parameters, designed for efficiency and faster inference with limited computational resources (e.g. DistilBERT)
- Model modality
  - Language Models: Understand, generate, and manipulate human language across various tasks (e.g. GPT-4; Gemini)
  - Text-to-Image Models: Generate images from textual descriptions (e.g. DALL-E2)
  - Text-to-Music Models: Generate music from text descriptions (e.g. OpenAI's MuseNet)
  - Text-to-Video Models: Generate video content from textual descriptions (e.g. TAAI)
  - Speech Synthesis Models: Generate human-like speech from text input (e.g. Google's Tacotron 2)
  - Image Captioning Models: Generate textual descriptions or captions for images (e.g. Google's Show and Tell)
  - Language Translation Models: Translate text from one language to another (e.g. Google Translate)
  - Question Answering Models: Provide answers to questions based on textual input (e.g. BERT)
  - Summarization Models: Create concise summaries of longer texts (e.g. BART)
- Model architectures
  - Generative Transformer Models (GTMs): These are a type of artificial neural network architecture particularly adept at handling sequential data like text
  - Generative Adversarial Networks (GANs): These involve two neural networks pitted against each other
  - Variational Autoencoders (VAEs): These use an encoder-decoder structure to learn latent representations of data
  - Autoregressive Models: Autoregressive models predict the next item in a sequence solely based on preceding items, employing Maximum Likelihood Estimation (MLE) for training
  - Convolutional Neural Networks (CNNs): Learning patterns from images (or sound, which can be translated into images) and generating new, realistic ones through techniques like transposed convolutions
  - Recurrent Neural Networks (RNNs): These are well-suited for sequential data like text or speech. They incorporate a concept of memory, allowing them to consider past information when processing new data points (e.g. Long Short-Term Memory networks)
- Model core training methods
  - Combined Loss Training: Leverages multiple loss functions simultaneously to optimize a model's training process (e.g. reconstruction loss and KL divergence loss for VAEs)
  - Adversarial Training: Two neural networks (generator/discriminator) are trained simultaneously but with opposing goals (e.g. used by GANs)
  - Supervised Learning: Equips models with the ability to learn patterns from labelled data, guiding them in creating new, similar outputs (e.g. used by CNNs)
  - Unsupervised Learning: Allows models to discover hidden structures and patterns within unlabelled data and is used by models not focused on strict replication of existing data (e.g. VAEs)
  - Reinforcement Learning: Empowers models to learn through trial and error, iteratively refining their creations based on rewards received for achieving desired outcomes (e.g. used by GANs with policy gradients)
  - Deep Learning: Using techniques like Maximum Likelihood Estimation, Backpropagation, and Self-Supervised Learning (e.g. e.g. used by GTMs)
  - Backpropagation Through Time (BPTT): unfolds a RNN during training, allowing error signals to flow back through time steps and improve the model's handling of sequential data

Note: This is work in progress!



Near-AGI = near Artificial General Intelligence

https://www.linkedin.com/pulse/how-generative-ai-has-supercharged-future-work

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Current Business Trends

- Types of AI Usage
  - Window-dressing
    - Basic digital projects branded as AI initiatives
  - Practical Applications
    - Customer service automation, reducing call handling times
    - Routine administrative tasks like data extraction and document drafting
  - High-Value Applications
    - AI aiding in professional tasks such as legal due diligence and financial research
- Challenges and Limitations
  - Inconsistent quality of AI-generated work requiring review
  - Data privacy concerns and lack of internal AI expertise hindering broader adoption
  - Technical reliability issues with early AI tools (e.g. autonomous driving vehicles)

Potential Applications to B2B

- Generating new business ideas
- AI-powered negotiation support
- Enhanced B2B customer service and support
- Generating realistic training scenarios for staff
- Visual content creation for personalised marketing
- Data management and security

The Economist: How Businesses are Using the AI

https://www.economist.com/business/2024/02/29/how-businesses-are-actually-using-generative-ai

- The article discusses the **cautious yet evolving adoption** of generative AI technologies like GPT-4 by businesses, highlighting the **significant financial enthusiasm** in the tech sector contrasted with the **still-limited and experimental use of AI in practical applications**.

- It explores how companies are integrating AI into various aspects of their operations, from low-level administrative tasks to high-value professional work, while **addressing the challenges and slow pace of widespread implementation**. Despite notable examples of productivity improvements, the article emphasises that **substantial economic benefits from AI are not yet fully realised**, and broader impact will require time and further technological refinement.

The Future

- Utopian Dreams about AI
  - AI Assistants for All
    - AI assistants seamlessly integrated into our lives, managing schedules, personalising healthcare, and providing real-time education
  - Supercharged Science
    - AI could accelerate scientific discovery, leading to breakthroughs in medicine, materials science, and clean energy
  - Enhanced Creativity
    - AI could collaborate with artists and designers, leading to new forms of artistic expression and innovation
  - Personalised Learning
    - AI tutors could tailor education to individual students' needs, maximising learning potential.
  - Robotic Workforce
    - Advanced robots could handle dangerous or repetitive tasks, freeing up humans for more creative and strategic work

The Future

- Dystopian Fears about AI
  - Fear of Being Replaced
    - This is the worry that AI will automate jobs and make human skills obsolete
  - Misinformation and Deepfakes
    - Generative AI could be used to create very realistic but fake videos or text, making it difficult to distinguish truth from fiction.
  - Loss of Control
    - As AI gets more complex, some fear we might not understand how it works or be able to control its decision-making

ChatGPT: These anxieties are valid, but it's important to remember that AI is still under development. There are researchers working on ways to ensure AI is used ethically and safely.

https://www.cinemagia.ro/filme/terminator-2-judgment-day-861/postere/4885323/#gallery

Generative AI Capabilities Timeline

	PRE-2020	2020	2022	2023?	2025?	2030?
TEXT	Spam detection Translation Basic Q&A	Basic copy writing First drafts	Longer form Second drafts	Vertical fine tuning gets good (scientific papers, etc)	Final drafts better than the human average	Final drafts better than professional writers
CODE	1-line auto-complete	Multi-line generation	Longer form Better accuracy	More languages More verticals	Text to product (draft)	Text to product (final), better than full-time developers
IMAGES			Art Logos Photography	Mock-ups (product design, architecture, etc.)	Final drafts (product design, architecture, etc.)	Final drafts better than professional artists, designers, photographers)
VIDEO / 3D / GAMING			First attempts at 3D/video models	Basic / first draft videos and 3D files	Second drafts	AI Roblox Video games and movies are personalized dreams

Large model availability: First attempts · Almost there · Ready for prime time

Bi (2023)

(My Personal) Conclusion

- Current state
  - Useful for mundane tasks
  - Useful for innovation
  - Useful for fun
  - Useful for criminals
  - ChatGPT can make mistakes; Gemini may display inaccurate info

- Some questions
  - Where are we going to be in 10 Years?
  - What would be breakthroughs in the B&IM field (what's on the wish list)?
  - Should we be worried?

References

- Bi, Q. (2023). Analysis of the Application of Generative AI in Business Management. Advances in Economics and Management Research, 6(1), 36-36.