VIII
MEDIA ETHICS
CONFERENCE
COIMBRA 2024

Inteligência Artificial como um caminho para a Inteligência Artificial Generativa

um percurso que salvaguarde a dimensão humana

Luis Borges Gouveia, Maria Beatriz Marques, Miguel Santos

lmbg@ufp.edu.pt, beatrizmarques35@gmail.com, miguelnuno@simbolodememoria.com

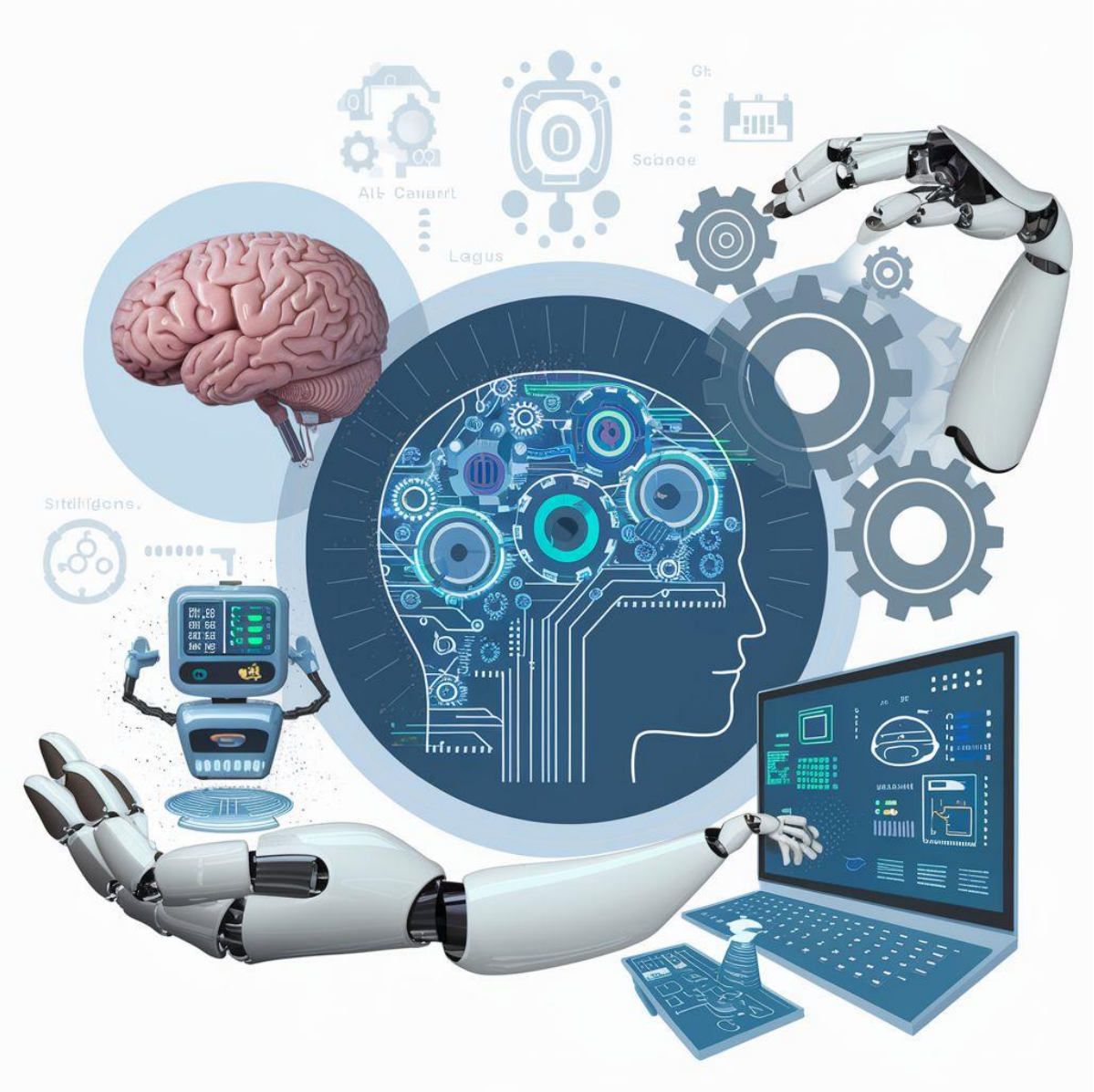
UFP – CITCEM, FLUC – CITCEM, FLUC

Inteligência Artificial como um caminho para a Inteligência Artificial Generativa – capacidades e serviços

- **Evolução da Inteligência Artificial (IA)** tem avançado de forma exponencial (parece estar a diminuir o ritmo), proporcionando novas e variadas aplicações. Desde logo, automatizando tarefas complexas, com impacto na forma como são tratados os dados e também nos processos de aprendizagem, raciocínio, decisão e planeamento, bem como o reconhecimento de conteúdos multimodo/multimédia;
- **IA Generativa (IA Gen)**: permitindo a criação de modelos que aprendem e interagem autonomamente com os utilizadores de sofisticação crescente;
- **IA Geral (IAG)**: como destino de uma IA que assegura a realização de atividades de base cognitiva de um modo geral, sem distinção do que é esperado de um ser humano, capaz de realizar a tarefa em causa constituindo-se como alternativa ao trabalho cognitivo de base humana;
- **Questões Éticas**: a IA Gen levanta **questões éticas** significativas e potenciais riscos que a sociedade precisa de abordar para evitar um uso e exploração descontrolado e com impactos não desejados, desde logo **percebendo e integrando o seu uso.**

Inteligência Artificial (IA)
Artificial Intelligence (AI)

- Stuart J. Russell & Peter Norvig (2019) definem inteligência artificial como um *conjunto de teorias e técnicas usadas para criar máquinas capazes de simular a inteligência humana.*
- Definição mais abrangente: área da ciência dos computadores que estuda a *criação de máquinas inteligentes que trabalham e reagem como os seres humanos, aprendendo, planeando, classificando, resolvendo problemas e reconhecendo dados, informação e conhecimento com o objetivo* ***de criação de aplicações autónomas ou de suporte à atividade humana*** (Gouveia, 2023).



Brij Kishore Pandey
@brijpandeyji

The AI Universe

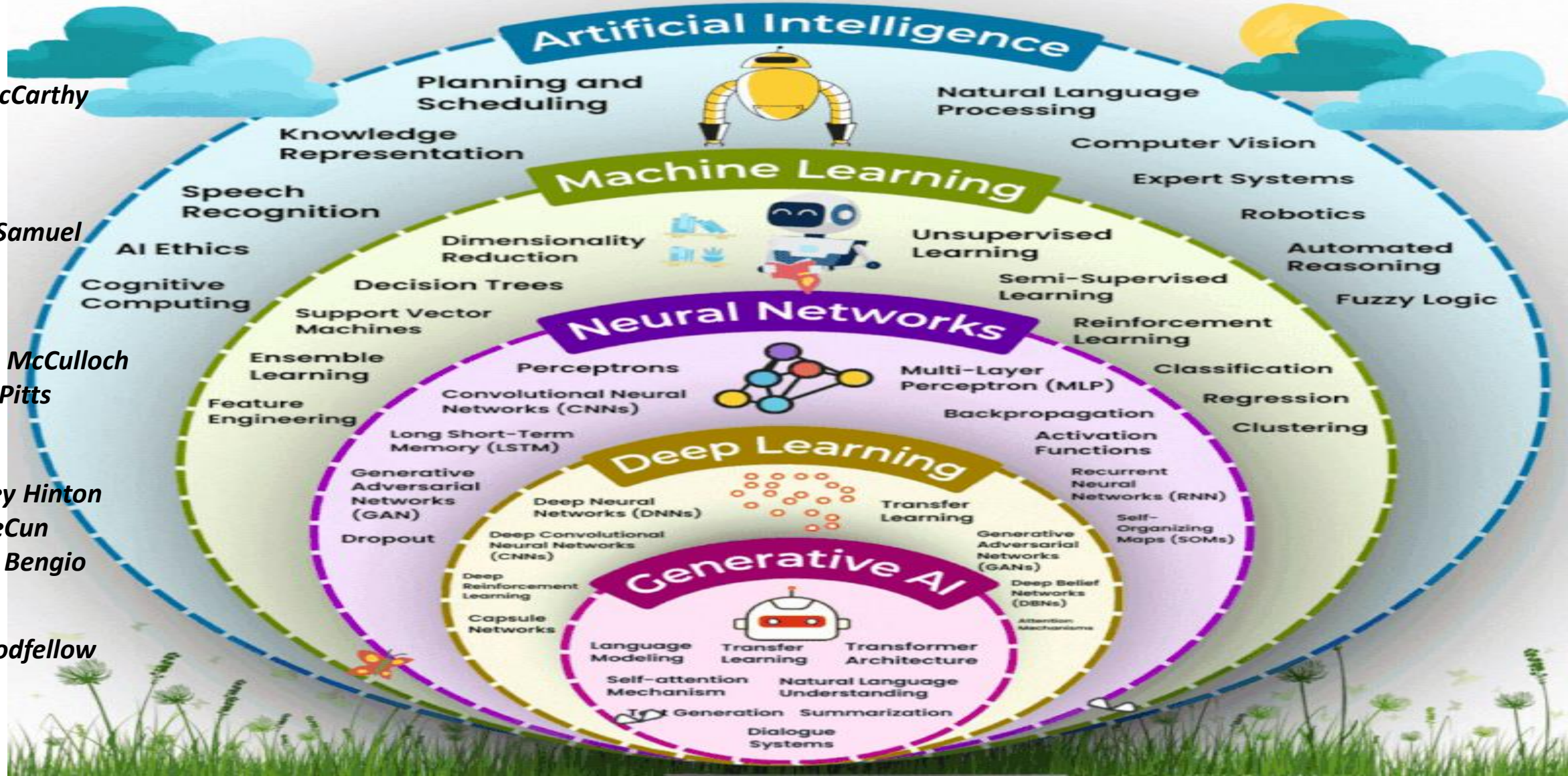
1956
John McCarthy

1959
Arthur Samuel

1943
Warren McCulloch
Walter Pitts

1990's
Geoffrey Hinton
Yann LeCun
Yoshua Bengio

2014
Ian Goodfellow



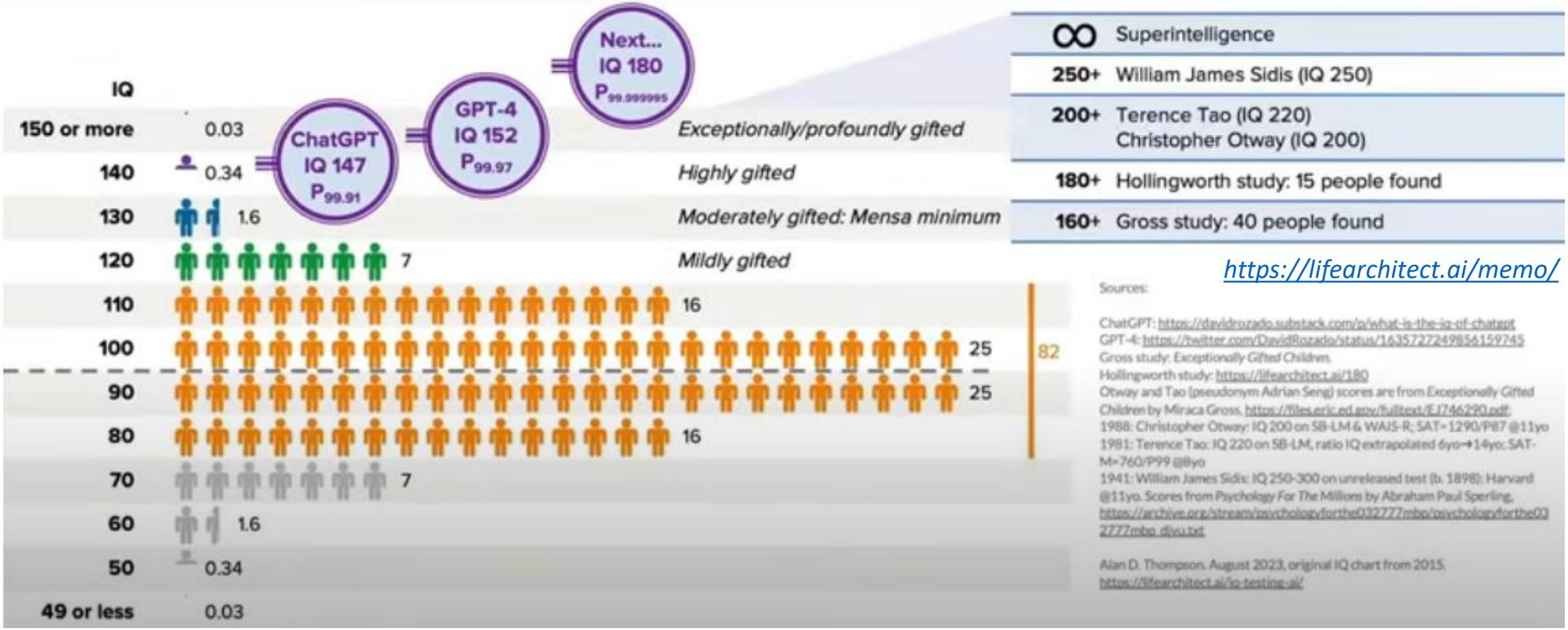
IA Generativa (IA Gen)

Generative AI (Gen AI)

- Uso de *modelos de aprendizagem profunda com capacidade de gerar texto, imagens e outros conteúdos de qualidade, baseados nos dados com que os modelos foram treinados;*
- Recorre a **redes neuronais** e estratégias de aprendizagem (supervisionada, não supervisionada e por reforço) para a elaboração de **modelos** que, pela sua escala e estratégias de equilíbrio na afinação dessas redes, permitem a **geração de conteúdos** pelo entendimento de questões ou pedidos de elaboração a partir de conteúdos;
- *Pela escala, medida em número de **parâmetros**, pela sua qualidade, função dos conteúdos (dados) de criação e treino desses modelos e nas estratégias de tratamento dos pedidos aos modelos, organizados em **tokens**, **emergem** capacidades para além dos dados utilizados na aprendizagem do modelo.*

IA Generativa (IA Gen)

Coeficiente de Inteligência Ser Humano vesus ChatGPT



Nota: o coeficiente de inteligência é apenas um parâmetro de comparação; não avalia a inteligência nas suas variadas dimensões…

Do digital aos modelos de linguagem

- A base de codificação **binária** da informação em computador originou o digital, que proporcionou a representação em multimodo e as funcionalidades de interoperabilidade, tornando os computadores instrumentos universais para processar, armazenar e comunicar informação;
- As redes neuronais e a IA Generativa proporcionaram o desenvolvimento de **modelos de linguagem** que permitiram a libertação do digital e da representação binária para um entendimento do mundo por via da linguagem, primeiro e, posteriormente, da imagem, do vídeo e do áudio, enquanto par do ser humano.

Riscos associados à IA
(O que pode impedir o seu uso?)

- **Confiança** (*trustability)*: como assegurar que é seguro, confiável e justo o seu uso;
- **Responsabilidade** (*liability*): qual a responsabilidade legal, qual o contexto e o quadro de regulamentação;
- **Segurança** (*security*): como monitorizar e evitar usos não autorizados ou indevidos;
- **Salvaguarda** (*safety*): como evitar descuidos do ser humano, atos não intencionais e falhas de equipamentos e infraestruturas;
- **Controlo** (*control*): o que acontece quando a IA assume um processo e se pretende transferir novamente o controlo para o ser humano, ou parar o processo.

Riscos associados com a IA Generativa

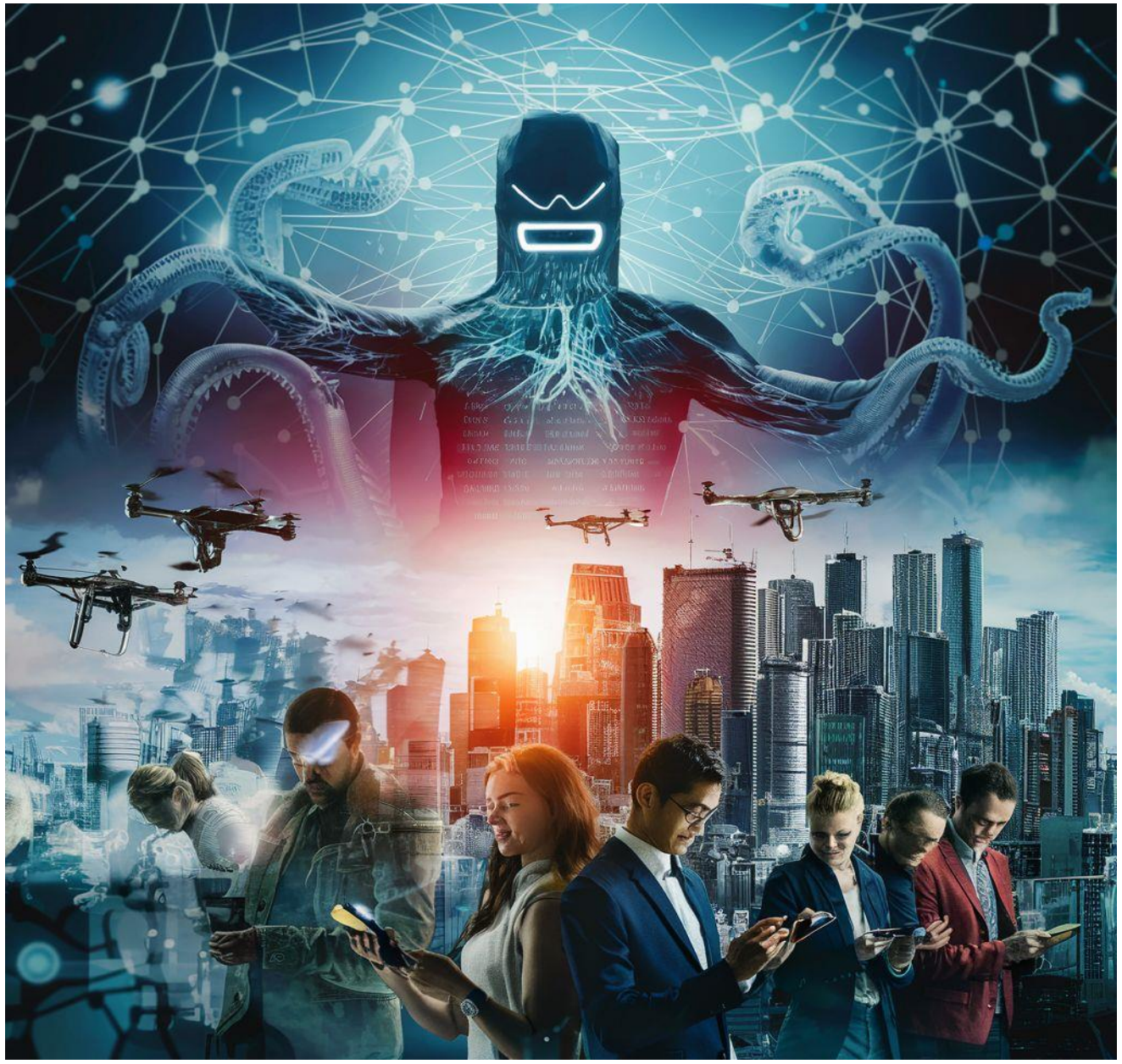
Adaptado de https://www.youtube.com/watch?v=9Rb9R7oTRks

- Violações;
- Privacidade de dados;
- Propriedade intelectual;
- Integridade académica;
- Resultantes do processo de geração dos modelos de dados;
- Viés que resulta dos dados de treino;
- Alucinação;
- Resultantes de uso não controlado;
- Falta de criatividade, ou de inovação, ou de ambos (tendencialmente, conteúdo de menor qualidade);
- A Web está a ficar tomada por conteúdo gerado de forma pela própria IA Generativa.

- **Uso responsável e ético** da IA
  - Uma estrutura a que as organizações recorrem para mitigar os riscos e desafios relacionados com o uso da IA, tanto de uma perspetiva ética quanto legal;
  - Definido por 5 princípios:
    - justiça;
    - fiabilidade e confiabilidade;
    - privacidade e segurança;
    - transparência; e imputabilidade.
- IA explicável (*Explainable* AI – XAI)
  - procedimentos e métodos que as organizações usam para compreender e confiar nos resultados gerados por algoritmos de aprendizagem máquina. De modo a melhorar a experiência do utilizador, dando a opção de verificar os resultados;
- *The Responsible Machine Learning Principles*: https://ethical.institute/principles.html

A relação com o algoritmo e a inteligência artificial

Ao tornar mais opaco o algoritmo, ao delegar na máquina o processo de decisão, também estamos a alterar o equilíbrio entre o ser humano e a máquina e a sua/nossa relação de poder.



"When most daily tasks are automated with AI, the relevance of thinking will be much more significant"

Kelwin Fernandes is the founder of NILG.AI

***Competências** para um contexto de coexistência com a IA de forma a garantir o aperfeiçoamento, para permitir a diferenciação e a integração com a IA, **de modo equilibrado, ético e sustentável**.*

Entender *versus* Compreender

- **Entender** (Dic. Priberam, *https://dicionario.priberam.org/entender*)
- 1. Apossar-se do sentido de (o que ouvimos ou lemos); 2. Ser de opinião, julgar, verbo intransitivo; 3. Ser entendedor; 4. Superintender.
- **Compreender** (Dic. Priberam, *https://dicionario.priberam.org/compreender*)
- 1. Abranger; 2. Encerrar; 3. Conter; 4. Entender; 5. Alcançar com a inteligência; 6. Perceber; 7. Notar; 8. Depreender; 9. Saber apreciar; 10. [Antigo] Achar (alguém) incurso em, ou culpado de; 11. Estar incluído ou contido.
- Capacidade dos modelos de IAG em compreender diálogos.

Questões e dilemas associados

- IA com controlo ou sem controlo?
- Incorporação de meios e capacidades em seres humanos;
- Incorporação de meios e capacidades em sistemas de armas;
- Autonomia em carros sem condutor, aviação comercial, setor da saúde, etc.

- Reflexão/questões:
  IA forte ou fraca? IA com controlo ou autónoma?
  Tal como no caso da segurança, quem guarda os guardas?

Planeamento
Explicabilidade
Alinhamento

- O ***AI SAFETY SUMMIT*** procurou criar as condições a nível global para lidar com estas questões:
  - The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023 *https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023*

Inteligência Artificial Geral (IAG)

(AGI – *Artificial General Intelligence*)

- IA com inteligência semelhante à humana: estes sistemas podem aprender, pensar e compreender conceitos em um nível semelhante ao dos humanos. Tem a capacidade de generalizar o conhecimento em vários domínios do conhecimento e de adaptação a novas situações, tornando a IAG versátil – é uma meta para o desenvolvimento futuro da IA (próximo? 2027-30?)
  - Flexível e capaz de realizar tarefas múltiplas e diversificadas;
  - Pode ser programada para aprender de forma contínua e ensinar-se a si própria.

As disciplinas principais da IA

Pág. 26 de Russell, Stuart & Norvig, Peter (2019). *Artificial Intelligence. A modern Approach*. (3rd Edition). Pearson (existe uma tradução Brasileira).

- Processamento de linguagem natural;
- Representação de conhecimento;
- Raciocínio automatizado;
- Aprendizagem máquina;
- Visão computacional;
- Robótica.

- Alan Turing concebeu um teste que permanece relevante até aos nossos dias. No entanto, resolver o teste de Turing, não é prioridade da IA, mas sim estudar os princípios básicos da inteligência:
  - **Racional**: O desafio do voo artificial teve sucesso quando os irmãos Wright e outros pioneiros pararam de imitar os pássaros e começaram a usar túneis de vento e aprender sobre aerodinâmica. Os textos de engenharia aeronáutica não definem como objetivo criar "máquinas que voem exatamente como pombos a ponto de poderem enganar até mesmo outros pombos";
  - Mas… GPT-4 foi considerado humano 54% das vezes, logo passando o teste, Jones, C. and Bergen, B. (2024). People cannot distinguish GPT-4 from a human in a Turing test. Arxiv, *https://arxiv.org/abs/2405.08007* .

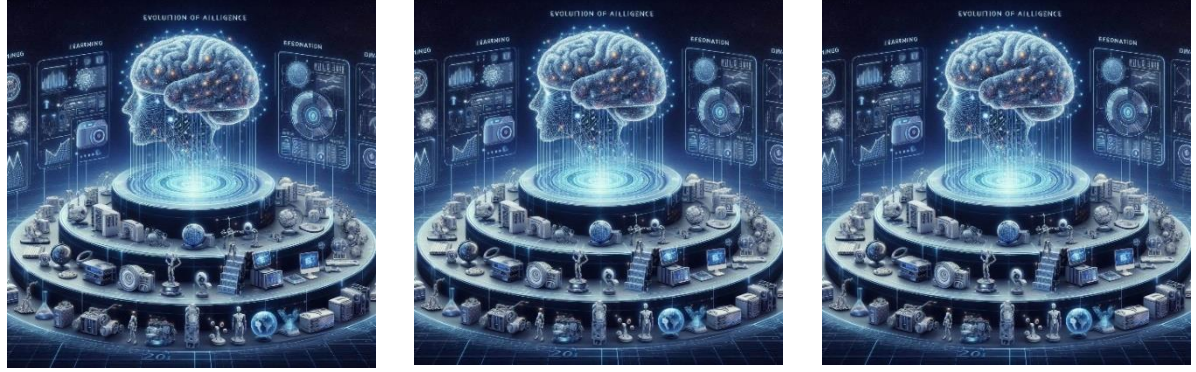
Pensar numa IA com valores e princípios é acima de tudo pensar numa Inteligência Artificial Responsável

- No caso de Portugal estes valores e princípios devem considerar:
  - A **Constituição Portuguesa**, nomeadamente:
    - A dignidade da pessoa humana;
    - Uma sociedade livre, justa e solidária.
  - A **Constituição Europeia**, nomeadamente:
    - Os direitos individuais;
    - As liberdades individuais.
  - A **Declaração Universal dos Direitos Humanos**, nomeadamente:
    - O direito à vida;
    - O direito à segurança.
- Guia para uma Inteligência Artificial ética, transparente e responsável na Administração Pública (PT, Gov)
  https://www.tic.gov.pt/documentos/guia-para-uma-inteligencia-artificial-etica-transparente-e-responsavel-na-administracao-publica

Lei da UE sobre IA (2024, processo legislativo iniciado em 2021). Níveis de risco

https://www.bloomberglinea.com.br/tech/europa-sai-na-frente-em-regulacao-para-ia-e-aponta-diferentes-niveis-de-riscos/

Comentários finais



- **Evolução da IA**: a **Inteligência Artificial Geral (IAG)**, independentemente da definição exata, a data de ocorrência está mais perto: variando de 2027 a 2030, como as datas mais próximas;
- **Automação e Serviços**: a IA Gen já possibilita a automação de tarefas complexas, alterando o potencial de automatização de processos de **aprendizagem, raciocínio, decisão e planeamento**;
- **Questões Éticas**: o avanço da IA, IA Gen e ainda mais a IAG, levanta **questões éticas** importantes, como o controle da evolução tecnológica e os riscos associados com implicações para indivíduos, grupos e sociedade em todos os setores de atividade humana;
- **Entendimento *versus* Compreensão**: sé apresentada a diferença entre **entender e compreender** e como os modelos de IAG podem responder a essas capacidades, pelo que urge **a separação clara do que é humano e do que é máquina, bem como assegurar mecanismos de salvaguarda adequados.**