# EU AI Act

Dave’s comments

6 March 2025

# Some Downsides of Current AI …

- Trustworthiness: AI systems have a tendency to make things up (hallucinate) when they seem like plausible guesses
- Bias can be hard to detect and correct, harming fairness and inclusivity
- Often easy to bypass protections
- Challenges around IPR and copyright
- Limited commonsense knowledge, e.g. about everyday causality
- Weak in respect to human empathy, intuition and ability to navigate complex moral issues inevitable in the real world
- Energy hungry – competing with other demands for electricity, e.g. cars

# AI is Evolving

- AI is usually defined as computer systems capable of performing tasks that typically require human intelligence
- That is of course anthropomorphic, and AI could take forms which are very unlike human intelligence
- Example: predict how proteins fold given their amino acid sequence
- Generative AI is new, and we can expect further breakthroughs
- *Best to avoid precise definitions that would be rapidly obsolete*

- The imminent era of sentient robots and *really* smart agents given further advances*
  - remembering and reflecting on past experiences, continual learning and reasoning for life-long skill acquisition
  - intelligent, wise, patient, cooperative, and exhibiting a high emotional IQ
- Huge implications for society
  - Manufacturing, services, education, law, healthcare, scientific research, etc.
  - Potential for boosting prosperity and decreasing inequality if we get it right
- The AI Act is just the beginning and further legislation is inevitable

* As explored in depth in the novels by Isaac Azimov

# Open Questions

- Generative AI uses statistical prediction to generate content that matches the training materials
- In other words, it blends human generated content rather than exhibiting genuine creativity
- It can solve University level exam papers, but it is unclear how much this exploits memorisation rather than reasoning
- Generative AI is good at handling the complexities of everyday knowledge
  - AI models lack transparency
- Symbolic approaches are good for transparency and interoperability
  - Valuable for neurosymbolic systems
- Symbolic AI and logic are impoverished, limiting their usefulness in the real world
  - Logic is unsuited to support a broad range of reasoning techniques, e.g. analogies
- We need a middle ground that deals with symbolic everyday knowledge that is uncertain, imprecise, context sensitive, incomplete, inconsistent and changing
  - Plausible Knowledge Notation (PKN) as a proof of concept (W3C Cognitive AI CG)

# Some Technical Considerations

- The ACT has a large number of requirements (113 Articles)
- What are the practical requirements for implementing all of these?
- Example:

  Article 53 states that the developer will draw up and make publicly available a sufficiently detailed summary about the content used for training of general-purpose AI models, according to a template provided by the AI Office
- A more extensive study is needed, and could perhaps be the focus of a CG

- Current AI requires huge training corpora, whilst humans learn with vastly less
  - This is a problem for specialist knowledge
- Challenges for run-time attribution of IPR, e.g. when generating images, music, video, etc.
  - How to recompense artists fairly?
- Fine tuning models to comply with human values
  - Safe, secure and trustworthy; non-discriminatory
  - Techniques include reinforcement learning with human feedback

# Implications

- Users must be made aware that they are interacting with an AI system
  - This could be part of the web page content presented to users
- Users should be able to determine what resources were used for training AI systems
  - Licensing IPR and copyright
    - Role for ODRL

- AI generated content must indicate that  it was generated by AI
  - This should be part of the content metadata, e.g. for media formats like images, music and video
  - For text, it should be included as part of the text itself
- Editing tools should retain this distinction
  - Role for W3C?  For PNG and SVG definitely.

# Implications

- How does the AI Act impact research and innovation (Academia, SMEs, Giant companies) ?

  Provision for running AI systems in sandboxes - how easy is it to arrange these, i.e. can this be done without a large expense that penalises academia, etc.

- What are the implications for other countries?
  - Divergence in regulatory frameworks for AI could create barriers to cooperation and increase development costs for AI applications

- Opportunities for W3C?
  - Neurosymbolic systems that combine the strengths of neural and symbolic approaches
    - Symbolic information supports transparency and interoperability
    - Role for W3C standards in respect to metadata, identity management, policy languages, credentials and knowledge representation
    - Standards for web interfaces, e.g. WebNN, AR/VR, intent-based models, etc.
- Societal Values
  - AI facilitates widespread monitoring and influencing, but will the EU rules on this be enforced or worked around?

# AI Agents and Open Ecosystems of Services

- AI agents are rapidly improving
- Current agents, such as ChatGPT, Gemini and DeepSeek, forget everything you told them in past sessions
- Next generation agents will remember for much longer periods
  - Neurosymbolic systems using retrieval augmented generation and personal databases
    - Imagine a combination of AI and SOLID
  - Further out: use of episodic memory
    - I am researching potential techniques in relation to continual learning and reasoning
- Potentially learning a great deal of highly sensitive personal information

- Personal agents will use what they know about you to help you with services, e.g. arranging a vacation or a doctor's appointment
  - AI is very good with imprecise requests and missing details, exploiting what makes sense in the context
- Lots of open questions such as
  - Where is the personal data held?
  - How much is shared with 3rd parties?
  - How to ensure open and fair ecosystems?
- Should W3C facilitate open debate over how this works?
  - Intent-based services*
  - Credentials as a basis for trust

* Intents facilitate accessibility

# Your thoughts?