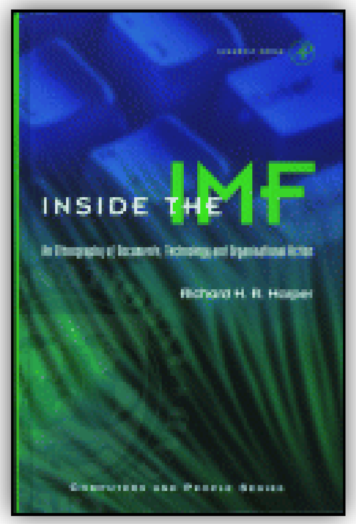
Intelligence and Ethics

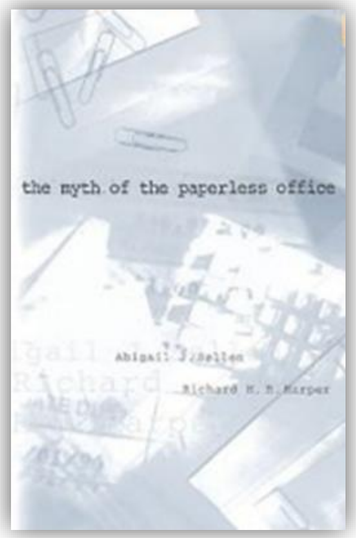
Professor Richard Harper

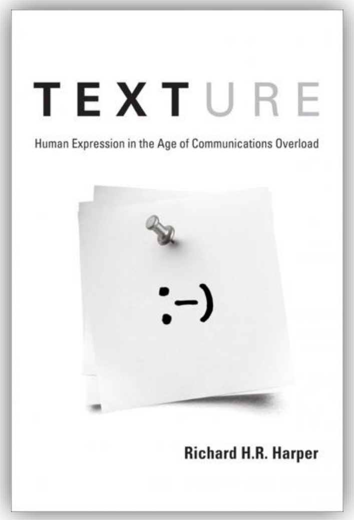
Overview

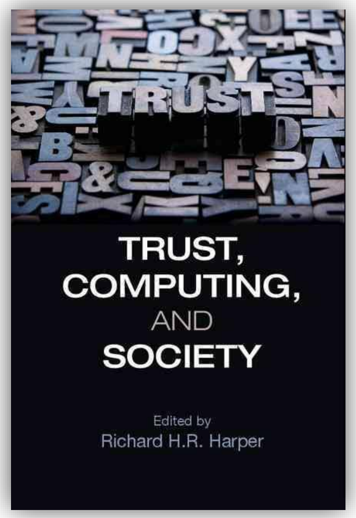
- Who am I?
- What is intelligence?
- A discussion of ChatGPT
- How it works, what it produces
- Is this technology intelligent?
- Is it ethical?
- Ways forward through good "HCI"
- The talk derives from a book I am currently writing called *The Shape of Thought* (McGill Press)

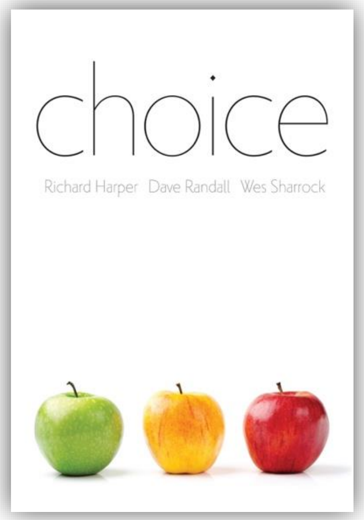
Who am I?: I have spent most of my career in corporate research - Xerox Euro-Parc, Microsoft Research, with start-ups and academe in-between

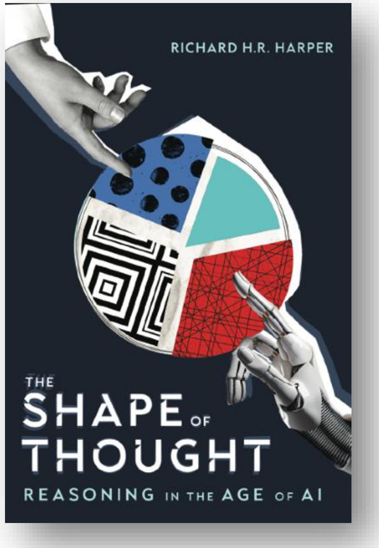

INSIDE THE IMF
An Ethnography of Documents, Technology and Organisational Action
Richard H. R. Harper
Computers and People Series











What is intelligence?

- Is it stuff you know?
- Or is it what you do with 'your' intelligence?
- Do we judge people by their intelligence?
  - Yes, in everyday life
- But in University? Why are you doing here?

Learning, intelligence, university

- At university, what do we do?
- Do we judge what people learn?
- So, we do 'rote' learning?
- Or do we judge the ability to judge given some learning?

- At university, we give some learning and then teach a person to judge with that learning

- So, this is what we mean by intelligence…(there are other definitions of intelligence, but this is the one we shall use today)
- Intelligence does not mean 'knowing stuff"
- Intelligence means the ability to judge given expertise, knowledge

ChatGPT

- What is it?
- A chatbot and an LLM – large language model
- Based on GPT-Number ‘x’
- Generative, Pre-trained, Transformer (GPT) of the LLM, currently version 4

- AI researchers (and the company selling it) say that ChatGPT it is intelligent
- So, does it judge?
- (that’s not the same as knowing stuff)

How does the LLM bit work?

- It’s a ‘model’ (which we could talk about some more) that represents the relations between words as likely frequency distances
- The most frequent sequence between one word and the next, such as raining – sky, or mud - earth make the distance between two words less, or in the opposite case, distant
- (are the words Arsenal 5, Man City 1, frequent? We shall come back to this example)
- The identified (‘modelled’) distances are added to other distances between instances (or tokens – this is what a word or part of a word ‘is’ in an LLM)
- For thousands of words/tokens – making millions of dimensions
- And then a great, complex, geometric system is produced
- Stored in tensors

The model: what is it a model of?

- There are lots of LLMs
- They all model patterns around the geometric distance between words when seen as a likelihood based on frequency in data corpora (the web)
- They are not representations of the world; they are representations of a calculated reinterpretation of things in the world – speech acts
- These geometries seem to be able to deliver outputs that reflect
  - Topics
  - Style
- All in terms of token patterns
- These outputs derive from user prompts

Knowing

- LLMs do not know what the patterns mean
- They do not know why words have the order they do, why they are ordered some way more frequently than in another, or what the words refer to
- Nor do not 'know' why one topic is associated with another, why the words, Mum and Dad, get associated with the word, Kids
- LLMs only know relationships between word frequencies, represented as geometric patterns in their tensors
- This is why I call them word geometry engines

Understanding

- What LLMs and Chat GPT know is:
- Not the meaning of something, but the patterned way ‘words’ are used, whatever those words may be

- Pattern not purpose
- Frequency not meaning

An example

- Arsenal 5, Man City 1
- Before last month, if you asked ChatGPT about these two teams, would this be the ‘word set’ offered back?
- This month, if you asked the same, would it be?

- Yes – but not because it is a fact, but because this phrase, Arsenal 5, Man City 1, is to be found on the web a lot
- Facts, truth, do no matter
- It’s the wisdom of crowds – what is being said often

Distance and tensors

- Or rather, this is a question of distance
- Tensors are the key concept or foundation of LLMs – this kind of AI
- Tensors are data stores of geometric space
- A vector through space
- A shape in space
- Or indeed any shape in multiple dimensions, (2, 3, 4, plus)
- Hence word geometries

Tensors

- The relations between tensors express relations between data types in terms of distance
- This distance can be conceived many ways
- Nearest points
- Common trajectory of a vector
- Similarity of bulk
- Etc, and so on

Tensors and meaning

- The important thing is that these are not 'truth functions'
- These distances are judged pragmatically
- I.e., the relevant distance is not the only distance, it is the distance that is seen to matter through frequency and likelihood
- These distances are selected as 'best fit' for the sought for output
- Feature engineers define these and work with the LLMs to produce or shape them
- What about 'guard rails'? (think of Arsenal/Man City)

Power of LLMs

- LLM machines are enormously powerful
- They combine billions of dimensions (distances)
- And they work fast!
- (and use huge amounts of energy)
- But they are not intelligent (in the way we have defined it)

Back to ChatGPT

- You enter a prompt
- A prompt is equal to a search query
- Depending on the prompt, the length, shape and substance of the output is determined
- E.g. What is the history of HCI in a paragraph, or more likely a list

Outputs

- What does the list entail?
- The most common definition
- When the most common is a word relation
- Then sub-definitions
- When the sub-definitions are the second most likely
- Cohered through a patterned style or written manner

Outputs can seem dazzling

The prompt: What is a LLM?

A large language model is *a type of artificial intelligence software that has been trained on a very large dataset of text in order to understand and generate language. These models can learn to predict the next word in a sequence, complete sentences, or even write paragraphs of text on a given topic. Some examples of large language models include GPT-3, which was trained on hundreds of billions of words, and OpenAI's Transformer XL, which was trained on 300 million words.*

Ok – dazzling outputs, but how is it to be used?

- What is ChatGPT4 doing?
- Is it judging or offering stuff?

- If it is intelligence
- What Is the judgement?

- If it's stuff,…
- Where does it get its stuff?
- Or is it providing 'answers to questions about stuff'? (judgement?)

What ‘it’ is (or might be) seems implied

- ChatGPT4 ‘acts’ as if it were a person – an agent that reflects as it answers
- It seems to judge
  - The text that is produced begins after a pause – as if some mind is considering
  - And the text is in the first person
- So is ChatGPT4 a kind of artificial person, an assistant who knows more than you do and can judge for you?
- Is it therefore intelligent?
- (In the way we defined intelligence at the outset)

The answer?

- No
- ChaptGPT only delivers likelihood, or better put, it delivers on likely distance between words
- It does not judge
- It cannot judge

So?

- Is it wrong to present ChatGPT as intelligent?
- Is it wrong to give its outputs the manner of an intelligent person?

Yes

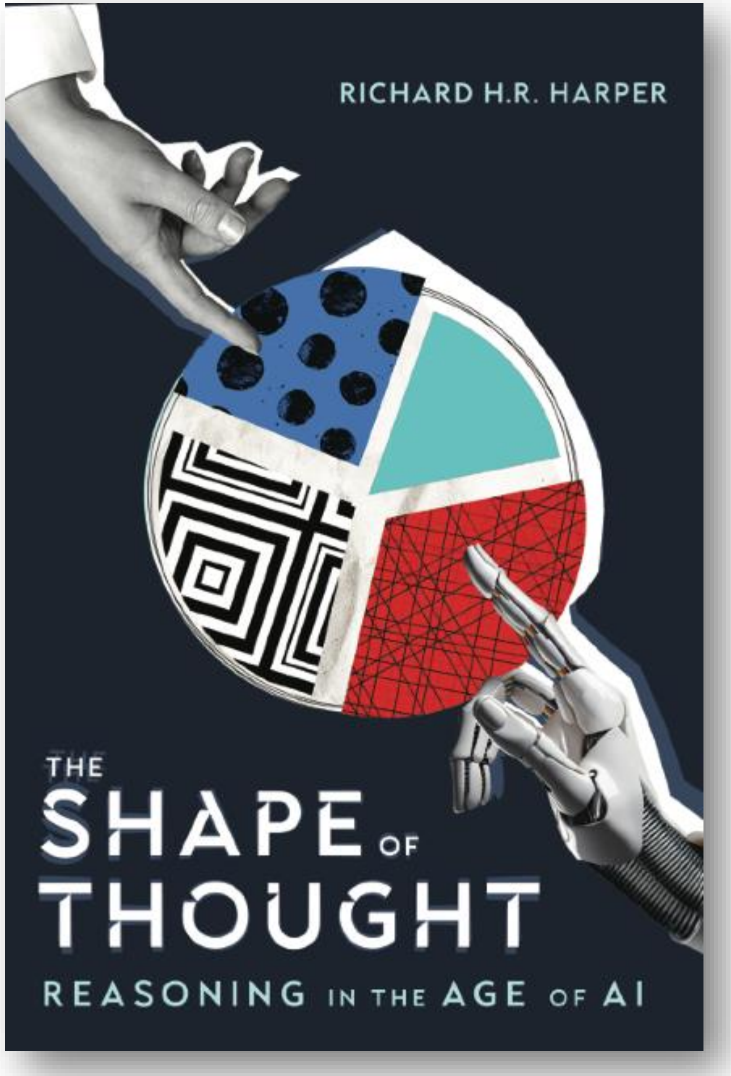
Chat GPT is unethical

Could it be made otherwise?

- Yes, with good HCI
- The design of the interface could make how it works and what it does visible
- It could remind the user that it only uses likelihood of words being used to answer questions posed in words
- It maps words to words
- It does not map words to facts

Further HCI

- It could also offer pointers to factual resources – Wikipedia entries, papers and blogs, etc
- It could be a portal between stuff and judgment, when judgement ends up being the user's problem
- ChatGPT4 could be a good tool
- But it needs good HCI
- It only has AI at the moment, and that is not enough


RICHARD H.R. HARPER
THE SHAPE OF THOUGHT
REASONING IN THE AGE OF AI

Some key texts for this talk

- Smith, D. C. (1982). The star interface: An overview. AFIPS’82 (pp. 515–552).
- Blackwell (2020). Objective functions:(In) humanity and inequity in artificial intelligence. HAU: Journal of Ethnographic Theory 9 (1): 137–146.
    - (see also Sarker, (2023). Enough With “Human-AI Collaboration”, ACM CHI).
- Basset et al, (2021) Ghost, Robots, Automatic Writing (Cambridge)
- Potts, (2023) The Near-death of the Author (Toronto UP).
- Shanahan, M (2024) ‘Talking about Large Language Models’, in Communications of the ACM, Feb, Vol 67. No. 2, pp68-79. https://doi.org/10.1145/3624724.
- Harper, The Shape of Thought, {McGill: forthcoming)