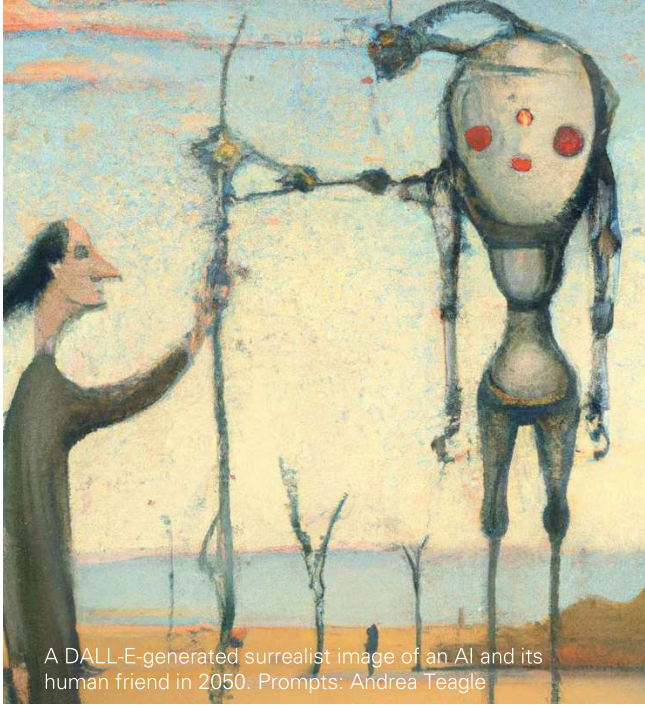
AI: the digital mirror

A new generation of artificial intelligence (AI) requires a new generation of policy-orientated research. While research agendas continue to tackle the thorny problems of previous-generation AI, recent advances in generative AI raise a new cohort of questions, ranging from the pragmatic to the philosophical. The HSRC's ***Dr Michael Gastrow*** *reflects on the research agenda created by this new wave of AI.*



A DALL-E-generated surrealist image of an AI and its human friend in 2050. Prompts: Andrea Teagle

Some revolutions happen quietly at first. In 2017, in a journal little known to social scientists or humanities scholars, a team of Google AI specialists published a paper titled 'Attention is all you need'. It introduced the transformer architecture, which revolutionised the performance of artificial intelligence (AI) algorithms and set in motion a shift towards a new generation of AIs, Generative Pre-trained Transformers (GPTs). Unlike older AIs, which served primarily as ranking functions (what show you might watch, the products you may buy, or which stocks might go up, for example), generative AI can produce novel text, images, sound, or video. The new AIs are 'trained' to better align their outputs with human values and interests.

Enabled by large investments in computing power and data scale, researchers at OpenAI, Google, and other technology leaders reached an inflection point beyond which AI was finally able to converse with humans in natural language. A cautious research and development process ensued, which took account of the serious risks inherent in powerful AI. Early commercial applications explored niche-use cases, for example, the 2021 release of DALL-E, which produces images from text prompts, and GitHub Copilot, which produces code from text prompts.

The technology exploded into public consciousness with the release of ChatGPT in late 2022, which applies the technology directly to text-based conversation with AI. This in turn led to a hype bubble. Science fiction nerds collectively shifted Artificial General Intelligence (AGI) to the non-fiction shelf, the investment world scrambled to monetise this breakthrough, and governments around the globe tried to catch up with the regulatory imperatives that the new technology creates.

What we should ask about AI

The most immediate questions are about benefit and risk. AI offers exciting new ways to learn, yet undermines traditional education systems. AI offers radical improvements in labour productivity, yet may lead to inequality, unemployment and upheaval. AI offers powerful new coding tools, yet also creates new cyber threats. AI offers remarkable new communication opportunities, yet can be used to mislead and manipulate. The impact is so broad that it is hard to imagine a sector that will be unaffected.

Some of the risk scenarios extend all the way to doomsday, speculating on the non-zero chance that AI could destroy humanity. However, unlike generations of science fiction movies, it would not be AI that destroys humanity, but rather humans that weaponise AI to do the job. The thought of authoritarian and malicious actors using AI for harm is, indeed, terrifying. Underlying all this is a shift in paradigm in which we let go of the notion of AI challenging humans for supremacy to embrace the idea that AI amplifies human capabilities and powers – and that in the end we will be faced with a scaled-up version of ourselves, both terrible and beautiful. AI does not give us anywhere to hide from ourselves: it is fundamentally an expression of humanity itself, of our collective ability to create machines in our own image.

The governance of AI is thus a pragmatic question of harnessing benefits and mitigating risks, as well as an existential question about the future of humanity. It is critical that social scientists and humanities scholars understand how the technology works, and that we develop scenarios for the technology's evolution and socio-economic impact.

The race to develop governance frameworks through legislation, regulation, and policy must be informed by sound research, including research that is grounded in the African context, and that reflects South Africa's developmental imperatives.

What AI asks of us

We also need to acknowledge that the mirror of AI asks questions of us, too. It asks questions of us as creative beings: if an artwork is created by a machine, based on text prompts from a human, and drawing on a silicon brain trained on the public internet, where does the creative spark lie? Is the resulting work an artefact of human culture

or cyborg culture? AI asks questions of us as conscious beings: human-like behaviour in machines is uncanny and, at the very least, raises questions about consciousness that philosophers have long explored only in theory, and can now test in practice. AI asks questions of us as humans: what would ultimately be left after machines take over much of our cognitive labour? What would we do? And, consequently, who would we be?

In this torrent of questions, it's important not to get carried away by the current. The limitations of generative AI are becoming increasingly apparent, and the counter-revolution has already begun. High-profile calls for a moratorium on AI development have entered the public discourse. Some people feel more comfortable doing their own thinking, rather than outsourcing parts of it to a machine with tentacles around the globe. AI-generated content remains imperfect, and can be incorrect and biased. The evident use of proprietary data for training raises questions of intellectual property rights, which are currently being tested in the legal arena. The initial hype of GPT has died down to an understanding that machines are obviously capable of comprehending language, and that our future is one in which we can converse with our fellow humans, or choose to interact with our simulacrum. It's likely that our children will take this for granted.

Researcher contact: Dr Michael Gastrow, a director of the Science in Society unit in the HSRC's Impact Centre
mgastrow@hsrc.ac.za

A DALL-E-generated image of a futuristic African city.
Prompts: Andrea Teagle