THE
ORANGE
PAPER
02

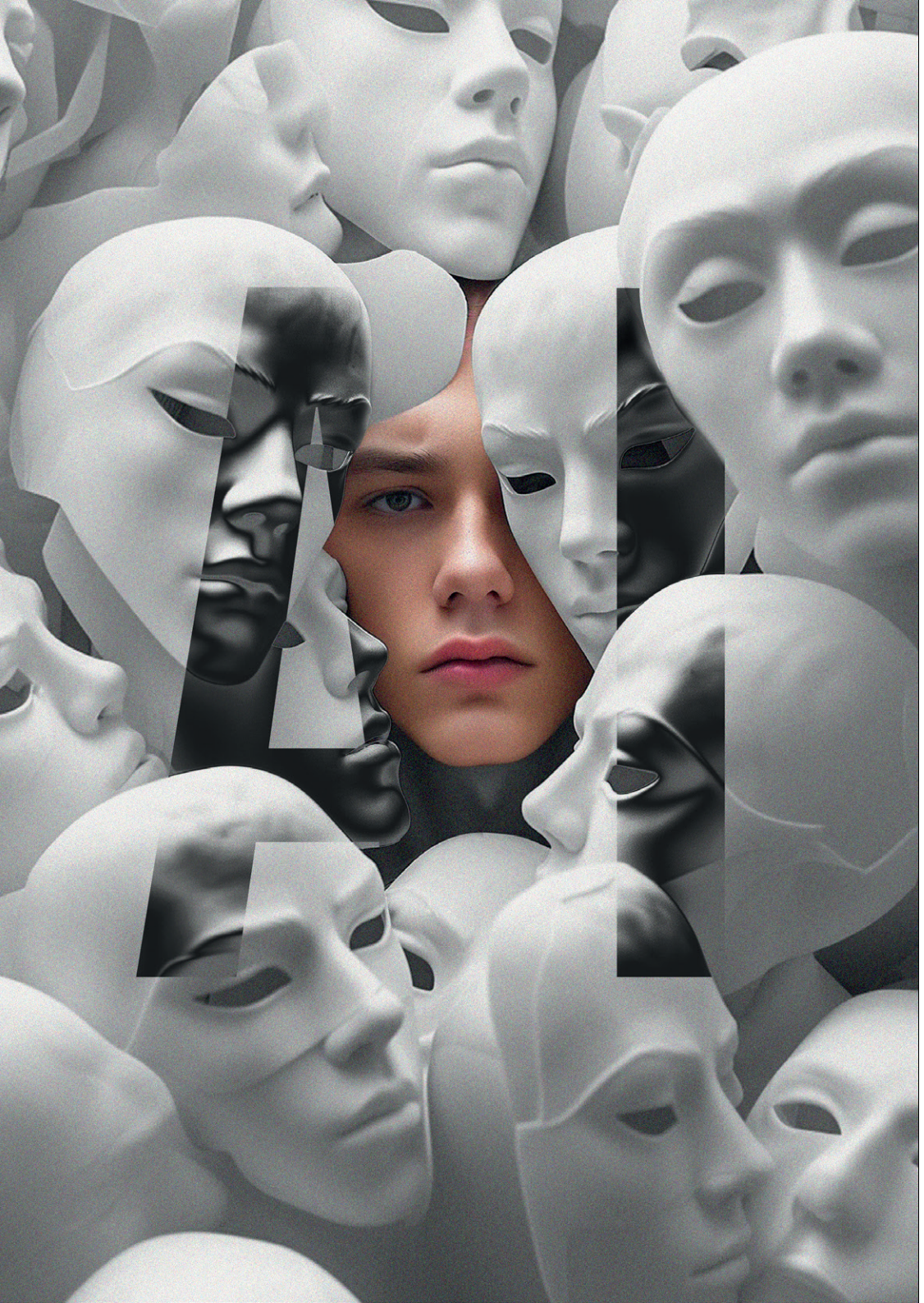
madano

THE
ORANGE
PAPER
02

madano

FOOL'S GOLD?

The unintended consequences of generative AI for communications



01. Introduction

‘MAY I HAVE YOUR ATTENTION PLEASE?’

There are immediate questions that need to be answered about how trust, truth and authenticity will be protected within a future of intertwined human and AI communications.

The launch of several groundbreaking content-generation and processing tools, such as ChatGPT, DALL-E3 and GitHub Copilot, have led to widespread discussion about the impact of rapidly evolving Artificial Intelligence (AI) on business and society. Available to the public and professionals alike, generative AI tools – which are often free or very low cost - are anticipated to transform many aspects of modern life.

So far, however, recent commentary on AI has been polarised around two major focal points.

At one end of the spectrum, generative AI's capacity to enhance human capabilities and increase productivity has been lauded. Despite its ease of access, generative AI arrived in the mainstream without a set of instructions – leading to the proliferation of guidance and opinion across media, from social to scientific, in all the flavours one might expect from online discourse.

Academic literature has considered the application of generative AI in science, healthcare and other high-value disciplines, such as computer programming. Much of the media discussion has focused on its (sometimes ethically questionable) uses or potential within corporate and learning environments. The communications industry, meanwhile, has been lit up by promises of immediate creativity and productivity gains.

At the other end of the scale, far-off future impacts of generative AI have also come to the fore. Some of these are tied to nefarious uses or applications within volatile and potentially destructive areas – such as biochemicals and warfare. The acceleration of AI capabilities has led others to speculate about how rapid advancement in agency, cognition and self-improvement could lead to Artificial General Intelligence; in which these learning models become, at first, indistinguishable from, and then vastly superior to, their frail and dispensable creators.

Yet there is a large chasm of our AI-infused future missing from the speculation – which is how AI is going to impact our existing systems of communications. Drawing on a range of voices and perspectives, we want to reset the conversation away from unbridled optimism. Instead, we offer a pragmatic look at the more nuanced and near-term evolutions and possible unintended consequences of generative AI. Here we find that there are immediate questions that need to be answered about how truth, trust and authenticity will be protected within a future of intertwined human and AI communications.

Generative AI arrived in the mainstream without a set of instructions - leading to the proliferation of guidance and opinion...in all the flavours one might expect from online discourse.

02. The evolution of AI

‘PRODUCTS OF THE INFORMATION AGE’

Generative AI is an innovation emerging from, and bound within, the structure of communications – in which information sharing and receiving has become not just social connection, but a form of leisure, status and monetisation.

Despite generative AI bursting into the public consciousness in 2022, AI and machine learning have been evolving for many decades. In fact, the growth of AI has been possible because of the late Information Age: the emergence of algorithm-led social communities, deep-pocketed tech firms, vast textual information stores such as Wikipedia, as well as spillover benefits in computing power from gaming.

In this regard, generative AI is an innovation emerging from, and bound within, the structure of communications – in which information sharing and receiving has now become not just social connection, but a form of leisure, status and monetisation. Ultimately, the critical near-term impact of AI will be to accelerate some of these trends, with major implications for businesses, governments and societies all over the world.

1950s-1960s

Early years Alan Turing published his ideas for evaluating machine intelligence - now known as the 'Turing Test'.

1950s - 1970s The birth of machine learning, natural language processing and neural networks. Significant optimism around potential for AI in the 1960s and early 1970s.

1964 ELIZA – the first chat bot was created at MIT, acting as a virtual 'psychotherapist' by responding to user input with preprogrammed responses.

1970s-1980s

1971 The Kenbak-1 launched – the world's first home computer.

1974 – 1980 The First 'AI Winter' – in which, following minimal progress in AI development and commercialisation, funding heavily declined.

Early – Mid 1980s Rise of formal logic and hand-programmed 'expert systems' as the leading approach to AI.

1980s Home computer ownerships expands with a range of products from Acorn, Apple and the BBC.

1980s-1990s

Late 1980s - early 1990s The Second 'AI Winter' – as limits of hand-programmed approaches emerge, optimism (and funding) around AI plummets again.

1990 Launch of Windows 3.0, a hugely popular operating system developed by Microsoft.

1998 The Internet – or World Wide Web – reaches 100 million users.

1990s-2000s

Mid 1990s-2000s Resurgence in interest in 'machine learning', an approach to AI in which computers are 'trained' through exposure to multiple examples, rather than programmed by hand.

1997 IBM's Deep Blue beats chess grandmaster Gary Kasparov.

2003 The Launch of MySpace.

2004 Launch of Facebook.

2006 Launch of Twitter.

2007 The first iPhone and the start of the smartphone revolution.

2010s

2012 AlexNet – a form of artificial neural network (ANN), won the major ImageNet computer vision competition – ushering in the boom in AI research and investment, particularly around deep learning.

2014 Generative Adversarial Network created – a deep-learning approach that 'creates' images by using two neural networks, in which one learns by the other evaluating its output.

2016 – DeepMind's AlphaGo beat Lee Sedol, a 9-dan professional, in Go.

2017 Seminal paper ('Attention Is All You Need') on a new deep-learning architecture now powering many generative models: the Transformer model.

2019 Pluribus becomes the first AI to beat professional players in six-player poker.

2010 The Arab Spring is viewed as the 'Twitter revolution' – facilitated as stemming from new forms of connection possible through social media.

2017 The Internet reaches 3.8bn users – over 50% of the planet's human population.

2018 Cambridge Analytica revealed to have manipulated elections across the world – including the UK's EU Referendum and the US Presidential election - through personalised advertising based on psychographic data.

2020s

2022 Public launch of ChatGPT by OpenAI – by January 2023 the chatbot becomes fastest-growing consumer software application in history.

2021 - 2022 Stable diffusion and DALL-E models break new ground in image-generation quality.

2023 Public release of GPT-4 by OpenAI – Microsoft Research declares the model "an early (yet still incomplete) version of an artificial general intelligence (AGI) system".[1]

2023 EU landmark AI legislation.

‘SCI-FI FUTURES AND AI BOOMTOWN’

Amongst this dramatic polarisation of AI endgames, reasonable assessment of more tangible risks and opportunities are getting lost.

Overhyped technology is a feature of a digital age defined by an exponentially increasing volume of content and exponentially decreasing attention spans. A range of technologies in the past two decades - The Internet of Things, Blockchain and NFTs – have been lauded by technologists, yet proven difficult for the public to understand and for companies, governments and societies to adopt.

It seems as if the stakes are even higher for AI. Google's CEO Sundar Pichai said in 2018 that "AI is probably the most important thing humanity has ever worked on", based on its ability to help solve some of the world's current problems in clean energy and climate change.

Technology as humanity's saviour is not a new concept - it's one that has defined successive waves of technological innovation since the Enlightenment: a European intellectual movement characterised by the celebration of scientific discovery and rational inquiry. Our language around technology is still imbued with centuries-old beliefs in the inevitability and desirability of scientific 'evolution' or, in its more positive encapsulation, 'progress'.

The Enlightenment conception of technological progress wasn't just fixing problems though; it was leading to a utopia in which problems were no more. The English philosopher and statesman Francis Bacon, in his incomplete novel, New Atlantis, suggested that developments in science and technology would ultimately lead to "the effecting of all things possible": thus, if humans can keep chipping away until they've uncovered all technological ideas, they will eventually create heaven.

It's an arresting vision of the future that speaks not only to the idea of a better life through technology, but also the inevitability of that happening.

Caution, fear, rhetoric and recrimination

Despite the technological optimism of Pichai and others, AI has also prompted strong caution. In March 2023, over 1,000 experts signed an open letter calling for a six-month pause on advanced AI development. Signatories included Apple co-Founder Steve Wozniak and revered AI pioneers such as Yoshua Bengio and Geoffrey Hinton. Other senior tech figures have also spoken out, warning that AI poses a threat greater than climate change, and that controlling it is one of our most pressing global challenges.

In response, Silicon Valley has rallied. In October 2023, US venture capitalist and software engineer Marc Andreessen countered these arguments in The Techno-Optimist Manifesto. Published on Andreessen's website, a16z.com, the Manifesto asserts that those calling for a slowdown in AI development are spreading "lies" about the future and encouraging people to be "angry, bitter, and resentful about technology". Echoing past utopians, Andreessen declares that "there is no material problem – whether created by nature or technology – that cannot be solved with more technology".[2]

Looking at the sun, but never in the mirror

Among this dramatic polarisation of AI endgames, reasonable assessment of more tangible risks and opportunities are getting lost. Ahead of 2023's AI Safety Summit, CEO of Cohere and co-author of the transformative 2017 paper, '*Attention Is All You Need*', Aidan Gomez, describes the doomsday scenarios linked to AI as a "distraction" from more real and relevant impact:

"This technology is already in a billion user products. That presents a whole host of new risks, none of which are existential... we should focus squarely on the pieces that are about to impact people or are actively impacting people."[3]

In a similar vein, US tech journalist Deborah Carver describes those on either side of the AI debate as "failing to understand what's actually going on". Blinded by their own beliefs, and deaf to the arguments of others, "all look at the sun", she says, "but never in the mirror".[4]

"This technology is already in a billion user products. That presents a whole host of new risks, none of which are existential... we should focus squarely on the pieces that are about to impact people or are actively impacting people." [3]

Aidan Gomez,
Cohere AI

04. The path for supercharged AI in ageing technology platforms

‘FRAGILE FOUNDATIONS AND STRUGGLING SYSTEMS’

Generative AI is emerging as both a potentially supercharged participant of the communications landscape as well as the next-generation saviour that could transform how we receive information and experience the internet entirely.

Our current communications landscape is a giant patchwork of privatised spaces and experiences gamified by algorithms pursuing attention. Internet 1.0 and 2.0 were initially defined by collective and open online experiences; but this has seen a significant shift in recent years, to a point where there appears to be much less understanding of what the internet really feels like for everyone else, or what motivates participation.

Generative AI is emerging as both a potentially supercharged actor in the communications landscape – flooding channels with deepfakes, SEO-manipulating articles and bot armies – as well as the next-generation saviour that could transform how we receive information and experience the internet entirely. Discussion and, critically, honesty about what communications really look like are, therefore, paramount in helping to understand the opportunities and risks for generative AI.

Decaying communications platforms

Dissatisfaction with the technology of internet communications has been steadily growing across corners of the web. Undoubtedly the side effects or shortcomings of technology platforms was already widely discussed: issues of platform failings (such as privacy) or the actions of bad actors (from bullies to governments). However, their position as the superior mediums of communications – from search to social connection – was barely challenged over the last twenty years.

In 2023, the technology magazine *Wired* printed a blog post by the author Cory Doctorow that described how the major platforms had succumbed to 'platform decay' – or as he pithily coined, 'enshittification': a growing feeling that digital products and platforms are deteriorating rapidly, to the detriment of their end-users.

Here, Doctorow suggests that the balance between user experience and revenue generation has become increasingly distorted to favour the latter:

"Here is how platforms die: first, they are good to their users; then they abuse their users to make things better for their business customers; finally, they abuse their business customers to claw back all the value for themselves."[5]

The argument follows that platform users are initially engaged through a valuable service –often delivered to them for 'free'. The monetisation of this for social media was based on advertising, and the levers of growth initially became the maximisation of time spent (more ads) or the ability to deliver better value to advertisers through a higher probability of conversion (better ads). Arguably, both of these have intrinsic risks anyway – as attention is suggested to be better captured through fear and hate, and targeted advertising is open for privacy abuse – however, advertising is often perceived as helpfully in-line with the user experience (such as 'sponsored results') or, at worst, a minor distraction.

The process of 'enshittification' suggests that, beyond this, a further twist of deterioration, is the perversion of the service itself: either directly by platforms seeking greater monetisation or through sellers competing outside of that monetisation in ways that diminish user

"Here is how platforms die: first, they are good to their users; then they abuse their users to make things better for their business customers; finally, they abuse their business customers to claw back all the value for themselves."[5]

Cory Doctorow

expectations. In Search, for example, users looking for a brand may as likely see advertisements for a competitor. In a 2023 blog, tech writer Ed Zitron describes Google as having "gone from a place where you find information to an increasingly manipulated labyrinth of SEO-optimised garbage shipped straight from the content factory".[6] Similarly, in an article in *The Atlantic*, Charlie Warzel declares that "Google Search, what many consider an indispensable tool of modern life, is dead or dying".[7]

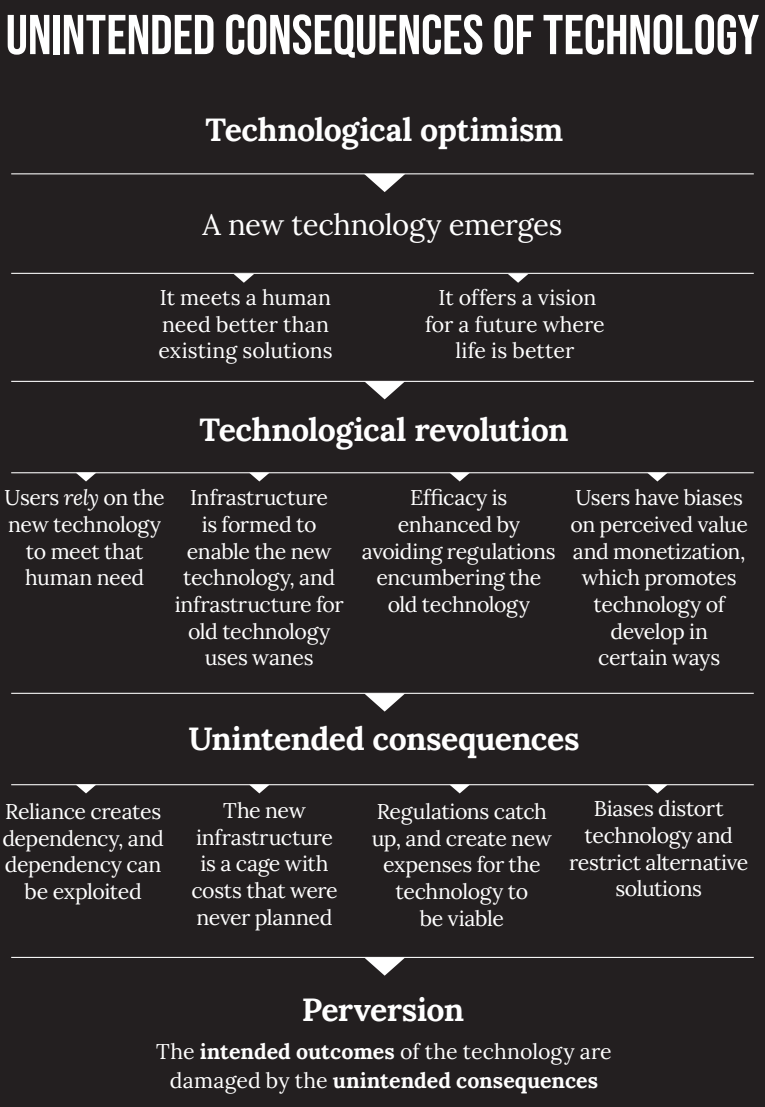
A pathway to unintended consequences

The potential contemporary failings of incumbent communications channels raise big questions about where generative AI might provide a solution to these shortcomings, serve to exacerbate them or carve its own trajectory. Even, however, if generative AI is able to develop a radical alternative to the dominant ways of the last 20 years for giving and receiving ideas and information – it may still inherit the same business model flaws that have led us here: owner expectations about growth and quasi-monopoly market dominance, and user expectations of accessing the latest technologies without direct payment.

The pathway to unintended consequences for a new technology, however, has a similar history to technology optimism – one that may just have not been written by the 'winners'. The most famous response to caution about the innovation of the Industrial Revolution, has become a deeply pejorative descriptor for stupidity and conservatism amidst the possibility of change: *Luddite*.

The issues are the still the same – that new technologies are initially cheaper or more readily valued to some extent because of failure to account for their indirect or future costs and can skirt regulation designed to prevent this from happening. It's only once embedded that these become a great issue; at which point the technology it replaced is a relic.

The perversion of a technology – the idea that Doctorow and others describe as the technology undermining that which it was intended for - is a more complicated path. But it's one that's been discussed in relation to a range of technologies. Automotives, for example, are a symbol of freedom of movement and ease of connection, despite evidence that suburban environments can limit residents' ability to walk and therefore interact in communities.



05. The real future of AI

‘CONTOURED AND CONTORTED COMMUNICATIONS’

The fundamental problem with unrestrained technology optimism is that it fails to acknowledge either the limitations of the existing systems that new technologies exist within or the well-known traps of unintended consequences.

Here we consider what the near-term pathways for AI might look like before we explore some of the emerging risks and issues that are likely to come to the fore in the near future – in ways that are broadly achievable with today’s technology.

WEB CONTENT & SEARCH

Hyper-personalised to individual tastes

Now

Currently, AI generates text that interprets organisations' owned content and reduces direct access – a prime example being Bing Chat. Sponsored content is integrated into AI chat as ad placements, while content is created by humans based on analysis of end-user engagement and an understanding of how best to hook people's attention – with attention valued as an end in itself, rather than a means to a more strategic goal.

Near Future

We expect to see an increasingly individualised online experience, with content (media, entertainment) being personalised to individual preferences and predilections at point of consumption. Independent online spaces will become limited, while machine-generated communications will grow. Human-created content areas, such as journalism, corporate reporting or media, may become heavily reduced as such content is replicated via AI, based on preferred user bias.

MARKETING & ADVERTISING

Heightened emotional manipulation

Now

With or without human review, hyper-personalised marketing content, linked to text, images and videos, is already a key component of online communications. Google's AI-powered ads create user-personalised content optimised to meet advertising performance goals. Users engage with content, such as social media, video or ads, which is fed by attention-oriented algorithms designed to trigger emotional reactions.

Near Future

Looking ahead, we anticipate a step-change in heightened emotional manipulation, with the blurring of lines between user-expected bias and neutrality touchpoints through the cross-channel integration of AI systems. Optimisation will also provide greater scope and potential for user manipulation. For example, advertising AI will increasingly access psychological data (derived from chatbot conversations) or physiological data (taken from wearable tech). But it will also be possible to manipulate users' emotional states through other AI-human interactions to achieve set objectives.

Tailored to individuals based on what they want to hear

Now

Today, users' interactions with organisations are mediated by AI in various ways, such as customer service, product or corporate information. Sometimes it is clear that the user is interacting with AI, whilst in other cases not. In some spaces interaction is possible with AI representations of 'real' humans, for example the range of celebrities – such as Snoop Dogg and Kendall Jenner – incorporated into Meta's chatbots.

Near Future

AI brand representations, through brand ambassadors or personifications, will continue to evolve. This process will be enabled by the elasticity of brand identity. A brand and its 'character' will become increasingly personalised, achieving broader appeal as communications are tailored to individuals based on what they want to hear.

AI systems being designed to identify partners and opportunities

Now

Information retrieval and synthesis are currently used to enhance knowledge and develop organisational points of view. Scaffolded AI-generated content, based on corporate viewpoints and tone of voice, is increasingly being used to communicate information.

Near Future

Stakeholder relations will become optimised through 'AI-to-AI engagement'. This process will involve AI systems being designed to identify partners and opportunities, or achieve consensus and cooperation on policy through negotiation, based on optimisation principles.

06. Truth, trust and intrusion

‘THE NEXT MISINFORMATION NIGHTMARE’

AI is likely to achieve greater access and proximity between companies and humans than ever before.

Undoubtedly many of the potential ways for generative AI to impact communications will be positive for end users – creating preferable ways of receiving information, faster information transfer and even previously unthought of ways to understand the world. Yet the risks are present, serious and deeply disturbing.

Personalization and trust

The first key challenge we envisage from generative AI is issues that might stem from personalisation.

Is content created by generative AI going to provide a genuinely fresh perspective? Or is it simply going to feed consumers what they want to encourage increased usage? Researchers have identified that models themselves may prefer to match responses to their expectations about user beliefs over more truthful answers – known as 'sycophancy' – without even being explicitly guided to.[8]

In the words of Ashley Gold and Sara Fischer, generative AI looks set to "trigger the next misinformation nightmare".[9] Generative AI models are now able to produce content so convincing it's becoming indistinguishable for everyday users from human-generated text or images. These capabilities have huge implications for the spread of misleading and false information. With factual content increasingly obscured and falsehoods harder to challenge, misinformation is becoming easier and easier to perpetuate. In 2024, for example, more than a third of US adults still believe the 2020 US election result was not legitimate.[10]

Hyper-personalisation exacerbates this, as misleading content can be tailored to prey on individual confirmation bias, causing certain beliefs and behaviours to become intensified and entrenched. If generative AI polarises opinion around contentious issues, it could result in more deeply entrenched social and political standpoints and a more fractious world.

Privatisation and truth

There is likely to be reduced spaces for truly in-public discussion in the years ahead, with increased privatisation happening through private AI-human interaction. Infinite user experiences will become available but will take place within 'walled gardens' and other controlled environments. It's still uncertain whether even those who control the specific systems will truly understand what is being communicated – as transformer models aren't readily interpretable and calls for greater interpretability are a lot quieter than those for faster and better performance.

Some commentators have questioned whether AI is capable of staying within predefined boundaries, or whether its cognitive capabilities could cause it to pursue its own goals, leading to human disempowerment, disutility and intellectual atrophy. Even if AI technology can indeed have embedded values to align to, there is a big question of whose values should they be. And technologists of different political perspectives have already started developing versions of AI based on this sentiment – such as Grok, funded by Elon Musk as a response to ChatGPT-perceived liberal bias.[11]

As generative AI begins to make decisions and achieve greater agency, the problem of value alignment becomes increasingly urgent. Could brand elasticity, for instance, lead to malleable and inconsistent values or even duplicity?

Generative AI looks set to "trigger the next misinformation nightmare". Generative AI models are now able to produce content so convincing it's becoming indistinguishable for everyday users from human-generated text or images.

Privacy and intrusion

Finally, there is the issue of intrusion into users' private lives. Harvard Professor, Bruce Schneier, writes:

"Near-term AIs will be controlled by corporations. Which will use them towards that profit-maximising goal. They won't be our friends. At best, they'll be useful services. More likely, they'll spy on us and try to manipulate us. This is nothing new. Surveillance is the business model of the internet."[12]

Certainly, over the next few years, AI is likely to achieve greater access and proximity between companies and humans than ever before. Whereas platforms currently gain insight into people's preferences and behaviours by leveraging emails, messages and searches, generative AI will build understanding through direct human interaction.

And herein lies the potential dependency risk: intimate knowledge of the end-user will make for an enhanced personalised experience but could ultimately expose us all to increasing levels of exploitation. In 2014, Facebook received public backlash for conducting research on how different information presented in news feeds would impact users' emotional states: with the answer being, a lot.

The world may still be approaching AI with naïve optimism about what might happen to its data, what is shared between platforms, and why: in the old internet, the insecurities that you tell your virtual therapists (e.g. search history, email data) may have been shared with advertisers. However, in the next frontier, advertisers could be paying to promote insecurities you never had before.

"Near-term AIs will be controlled by corporations. Which will use them towards that profit-maximising goal. They won't be our friends. At best, they'll be useful services. More likely, they'll spy on us and try to manipulate us. This is nothing new. Surveillance is the business model of the internet."[12]

Bruce Schneier,
Harvard Kennedy School

Conclusion

It's difficult to really know how this balance between risks and benefits will play out – as well as going against the grain of optimism and excitement. Like many digital technologies that have come before it, there's playfulness, intrigue and a sense of the limitless to generative AI; but it's a shine that fades if you have to think about the ethical implications of it every time you want to get an answer.

Fundamental to the way that Madano views its own role, and the role of the industry, in supporting emerging technologies is to thoughtfully shape the future – not just aimlessly cheerleading into it. It's incumbent on us all within the industry to upskill, be more knowledgeable; and ensure not only effective and efficient use of AI, but protection from the erosion of truth, trust and authenticity.

MADANO AND AI

At Madano, we've been using AI for many years. Our Insights Practice use machine learning models to enable us to gain insight into attitudes and behaviours at superhuman scale. We use AI to support our creative work and manage content and information flows. And we're constantly looking at where AI can be practically and meaningfully deployed to enhance our service delivery and improve outcomes for clients.

Leveraging our capabilities in data, analytics and technology, we also advise our clients on opportunities to embrace AI in ways that are measured and impactful.

Our approach to AI, as in all things, is people-first. We believe a balanced view and clear-eyed assessment of the costs and benefits can only come from an understanding that technologies are developed and deployed within the context of human businesses and human societies.

It is by understanding both the human and technological, people plus machines, and how each influences the other, that we help our clients face the unknown and shape their own futures.

References

1. Bubeck, Sébastien et al., 'Sparks of Artificial General Intelligence: Early experiments with GPT-4', 22 March 2023, arXiv:2303.12712
2. Marc Andreessen, 'The Techno-Optimist Manifesto', October 2023, https://a16z.com/
3. Quoted in Dan Milmo, 'AI doomsday warnings a distraction from the danger it already poses, warns expert', *The Guardian*, 2023, https://www.theguardian.com/technology/2023/oct/29/ai-doomsday-warnings-a-distraction-from-the-danger-it-already-poses-warns-expert
4. Deborah Carver, 'When the discourse misses the point: Techno-optimists vs. legacy media', *The Content Technologist*, 2023, https://www.content-technologist.com/rethinking-the-digital-discourse/
5. Doctorow, 'The 'Enshittification' of Tik Tok', *Wired*, January 2023, https://www.wired.com/story/tiktok-platforms-cory-doctorow/
6. Ed Zitron, 'The Rot Economy', *Where's Your Ed At*, 2023, https://wheresyoured.at/p/the-rot-economy
7. Charlie Warzel, 'The Open Secret of Google Search', *The Atlantic*, 2022, https://www.theatlantic.com/ideas/archive/2022/06/google-search-algorithm-internet/661325/
8. Mrinank Sharma et al, 'Towards Understanding Sycophancy in Language Models', 2023, https://arxiv.org/abs/2310.13548
9. Ashley Gold and Sarah Fisher, 'Chatbots trigger next misinformation nightmare ', *Axios*, 2023, https://www.axios.com/2023/02/21/chatbots-misinformation-nightmare-chatgpt-ai
10. Rachel Weiner et al, 'Republican loyalty to Trump, rioters climbs in 3 years after Jan. 6 attack', *The Washington Post*, 2023, https://www.washingtonpost.com/dc-md-va/2024/01/02/jan-6-poll-post-trump/
11. Kelsey Vlamis, Elon Musk vows to change his AI chatbot after it apparently expressed similar left-wing political views as ChatGPT ', *Business Insider*, 2023, https://www.businessinsider.com/elon-musk-vows-to-make-xai-chatbot-grok-politically-neutral-2023-12
12. Bruce Schneier (2023), 'AI and Trust' [https://www.belfercenter.org/publication/ai-and-trust]