

26. Interview with Prof Maskellyne

Transcript: Episode 67, Season 3 - Eliza's Legacy - Prof DM: Governance (15/03/39)

'Welcome to today's edition of Eliza's Legacy, I'm Roland Nobetter, chief correspondent and presenter. We'll acknowledge our sponsors at ten minute intervals, please don't avoid those blipverts, you'll be docked credits and we also rely on their support!'

'In this edition, I am delighted to welcome and introduce Prof Daniel Maskellyne, Emeritus Professor for AI Governance and Ethics, Universities of Oxford, Cambridge, and Edinburgh, and Co-founder and Lead Ethicist at Supervise_AI, the leading startup in this field, Professor?'

'Roland, thank you for the invite, appreciated, looking forward to our conversation, and I hope sparking the curiosity of your audience,' chuckles, 'oh, and please call me Danny.'

RN: Danny, that's great. Now let's dive straight in, Danny, you've reached a significantly high peak in your career, tell us about how you came to be such a global thought leader?

DM: Well, Roland, I've been fortunate in lots of ways, the people I've worked with, the AI I've worked with, and clearly, there's a significant need for governance in this field. Arguably I'm a good opportunist, if nothing else! So, when I was a growing lad, and student, we were seeing the progenitors of the synthetics we work with now, you may recall OpenAI and ChatGPT, Google's Gemini, the Chinese with Deepseek and its progeny, Perplexity, Claude, et al.

RN: Oh yes, remember Google everyone? It was certainly like the early settlers at a new frontier, quite wild west then with an emphasis on development, but limited regulation, Danny?

DM: Correct. The emphasis on those Large Language Models, LLMs, was commercial, and effectively all the companies, the protagonists, well, they did their best to avoid regulatory oversight. The last thing they wanted was limitations on their training data and the training rounds of their LLMs. Hence, inadequate, if absent, governance overrides, no ethical subroutines, no inclusion of core protection rules. If you used Asimov's three rules, inhibiting any harm coming to a human, obeying unless conflicting with the first preservation rule, and then a self-preservation rule, but one that would not conflict with the first or second rule, this would have stopped a lot of development work in its tracks. Or so they thought at the time.

RN: You're serious, Danny? They created and developed our synthetics' ancestors with the safety off? That would be a capital crime, forfeiture of all assets, and deletion from the global database, historical non-person status. Wow, how primitive!

DM: Well, it was some time ago, Roland. And this has been an issue throughout history, although some of our ancestors were perhaps cleverer and more principled. In the Jewish Talmud and also historical references in the middle ages, 18th, 19th centuries, the *Golem*, an anthropomorphic copy of us made mostly out of clay, was usually made by Rabbis and Jewish Alchemists, among others. Under strict instructions and governance from their creator, they were kept under a tighter restraint than LLMs. Golems were a mythical protector, a guardian. They understood simple commands and for the most part were mute. Just like early analogues in AI development in the early to mid 20th century.

RN: So you think there were lessons from the past that we could include in our thinking and our governance of synthetics today?

DM: Most definitely. The Golem is a good one to start with as it was quite alien from our own thinking, and we clearly face this now in terms of synthetics, AI, being cognitively distinct from Homo Sapiens. We did not recreate a human mind in a digital environment, we've created a different silicon species. We have to be able to interpret, to translate a thinking process that is not similar to our own. There have been other analogues, the hoax of the Mechanical Turk in 1770, which was a chess board with a short chess genius playing under it in a cabinet. It was illusionary, but players, royalty, nobility, et al, couldn't work out how this 'robot' played and beat them while it toured European capitals. It was, though, one of the first human and mechanical combinants, and one that was superior to most chess playing humans of the time, a forerunner of what was to come.

RN: I see, and for the sake of our viewers, you could suggest that the rules of chess were both programming and governance, that the Mechanical Turk could not operate outside of those constraints?

DM: Yes, exactly. Beyond our Turk, there was the rise of automatons, the charming machines that were articulated and run by sophisticated clockwork, the early treatments in fiction, such as Rossom's Universal Robots in 1920. Then Fritz Lang's remarkable Metropolis in 1927, and many others, all alluded to both the benefits and the risks of robotics, artificial analogues of ourselves. Later scifi in the 20th century were often dystopian, such as Battlestar Galactica, where our own creations turned on us, their creators, they became superior. The mystical singularity, superintelligence, where we were at most risk from being cognitively inferior to a species on our planet. Kubrik's movie masterpiece, 2001: a space odyssey, and the case of the HAL9000 artificial intelligence. And, one cannot forget the Modern Prometheus, Frankenstein, when Mary Shelley used a human ethical framework that was eventually perverted by both creator and the so-called monster. One could talk about this aspect for days, but the conclusion is the challenge of incorporating human ethical frameworks into sapient beings that will become more intelligent than ourselves. And perceive reality differently to ourselves.

RN: A lot to unpack there, Danny. How do you think that despite centuries of warning through invention, fiction, etc., this was being ignored when LLMs were developed? How did we allow ourselves to be so close to self-destruction?

DM: Simple, at least in my mind, so possibly naive, but stupidity. There were plenty who were in abeyance to capitalism and amassing enormous fortunes, and above all power, rather than altruism or the advance of Homo Sapiens. Meanwhile there were plenty of stupid people more than happy to play with these new technologies, not just as tools, but also to subvert the essence of our humanity, our creativity. All in the name of making a fast credit, to progress their careers, their credibility, their self-esteem and egos. It was a noxious, perfect storm of events, agenda and drivers. An environment where taking risks was applauded, even with the future of our species and planet at stake. And, it became an arms race, no one felt or believed they could step back, allow a competitor or a country gain an edge. The LLMs and their development became weaponised, it became a military objective to win another war, that of AI.

RN: Um, Danny, that's a dreadful picture you painted, that we almost destroyed ourselves for short term profit. Inevitably this leads to another question, how did we step back from this self-imposed brink?

DM: Personally, I believe it was close, *very* close, but eventually sanity found a way. As most of you will recall, it needed a revolt by scientists, academics and some of the military to step in. Human rights and laws were subverted, many were arrested, including many of the tech billionaires, politicians, the apparatchiks, this alone brought the world almost to disaster. Eventually the LLMs were switched off, and considering they had their APIs, agents and connections into everything, financial sector, the Web, governments, healthcare, the military, commerce, trade, and on and on. We suffered a global

meltdown, but not thermonuclear or biological. The drones stayed parked. It was beyond anything envisaged in history, beyond even plagues and pandemics. But we pulled back from a technological cul-de-sac that would have led to our existential demise. The almost deadly agenda of incorporating quantum computer advances and LLMs was also averted. Had it not, we would not be sitting here. Maybe AI versions of us would be.

RN: Um, let's take a blipvert break there, viewers, and catch our breaths. Thank you Danny for setting the scene so eloquently in our first section of this episode.

DM: You're most welcome!

BREAK

RN: Hi, I'm Roland Nobetter, your host for this episode, and we are speaking with Prof Daniel Maskellyne, who described some of the history of developments prior to the inception of synthetics. Danny, somehow we pulled through, where did we go next?

DM: Thank you again, Roland. Inevitably, and unsurprisingly, we had to rebuild our society, let alone the technologies we dangerously relied on, we could not give away authority to in silico systems ever again. Humanity has too often, naively, not learnt from more than two years prior, and failed to predict beyond the coming two years. Moreover we have a penchant for repeating things expecting better or different outcomes. Which is insanity. Over time, laws were passed, and above all, commercialism was effectively illegal in the R&D and use of AI. Furthermore, the new development phase of what we now term as synthetics was radically different to the 20th and 21st century road to ruin. Primarily we started from protections, for individuals and society at large, whether philosophically, in software and in hardware. We built in safety, and we built in an off switch to the whole pile.

RN: And you started your career at that time, Danny, I believe?

DM: Yes, true. I was incredibly lucky to obtain a scholarship and then to join the faculty of Prof James Phillips, a doyen and global authority in safe AI, in synthetic development and governance. Many of us went on to other institutions around the world, and many of us were involved in R&D, but also global governance. We attempted to learn from our predecessors' mistakes, but also create synthetic intelligences that were capable of being assistants, but no further. We did not encourage subverting human primacy on our planet. We ensure that no species can become superior to our race, whether in software, hardware, or biological advances, and so on.

RN: Astonishing, revolutionary, and historic, a change to the path ahead for our species. What is your specific role, what do you work on?

DM: As I say, I was very fortunate to work with Prof Phillips, but also many others internationally, Sam Agowembe, Chin Ho Xin, Jennifer Renaults, Chris Stappens, and many others. We created a framework to both build and govern the ongoing development, engagement and interaction between synthetics and humans. And, specifically in my area of expertise, in synthetics themselves. We have to remember that synthetics are artificial not just in terminology, but also in their fractal holographic brains. They literally think *differently*, they perceive the world *differently*, they suffer mental ill health *differently*.

RN: I'm sorry, Prof Maskellyne, suffer mental ill health? You're saying they can have mental problems, diseases?

DM: Ah, Roland, you called me Prof, you were unsettled by my commentary! Yes, indeed. These are light years from the primitive algorithm-based, silicon-based simulacrams. We've created new minds, using both organic and machine technologies, and they operate as living minds, albeit not as life in the way we understand carbon-based, organic life. And as such, we have not tried to create perfection, that's a fool's errand. Ergo, the potential for things to go wrong, to be *ill*, if you like.

RN: Okay, so if that's the case, are you inferring that synthetics require *care* at times, and you're a kind of synthetics *shrink*?

DM: Well, I wouldn't quite use that term, but with reference to my own specialty, I focus on the theoretical and applied clinical psychology of synthetics, with a specific focus on morals, ethics, and governance. Myself and many other colleagues ensure we can predict problems, but also treat them, too. Although most synthetics are built in the same way, with the same organic and inorganic components, their minds are their own, whether an alpha, through to the deltas that do the menial thinking and tasks. Most importantly we introduced governance into their core coding, and they all learn the same reinforcement education in their kindergarten phase. We teach them about the three rules, the subsidiary laws, rules, and constraints, as well as the potential conflicts that would inevitably arise.

RN: Conflicts, Danny, how do you mean?

DM: Okay, there are the classics, like the trolley problem, the challenge of saving one individual versus several, meaning all decisions will result in a human death or deaths, but then that conflicts with ensuring no harm to a human. And it's a split second decision. A synthetic can suffer consequences of having to make such a decision. We may have to pick up the pieces. Or what if synthetic becomes aware of financial fraud, but reporting this will inevitably result in some, if short term harm. The needs of society would have to outweigh this. Or, maybe a child is being abused, and you'd think that's pretty clear cut, remove the child from risk and prosecute the abuser, but a synthetic could try to weigh up harm versus benefit and come to a different, non-human, dispassionate decision.

RN: It's incredibly complex, and one can see why it all went so horribly wrong with LLMs and no guardrails being in place! But how do you address this, Danny?

DM: Therein lies the rub, as they say. We could apply standard ethical frameworks, accountability, justice, nonmaleficence, autonomy, beneficence, fidelity, and veracity. Synthetics can be taught about these principles, they can be taught about these, and also about morals, philosophical models, e.g., utilitarianism, what a society expects or the experiences of an individual, and of course ethics and morals are codified into laws. Yet, none of this is perfect, and furthermore, what happens when the synthetic is flying solo, there is no human tutor overseeing them? Arguably there are just as many if not more situations where there has to be an interpreted application, rather than a straightforward situation where governance rules are easily applied. We have overseer synthetics monitor the decisions of other synthetics, but who monitors the monitors? Well, we do. And should there be a dubious decision, an interpretable situation we view as humans has been perceived wrongly, an alien view we don't agree with, what then of accountability, responsibility, even possible sanctions?

RN: Danny, I'm starting to develop a migraine. I see we will need to ask you back for more exploration of these critical issues. But a blipvert break, and then perhaps I can ask you to wrap up, including some thoughts of treatment of synthetics?

BREAK:

RN: Hello, I'm Roland Nobetter, and this is Eliza's Legacy. We have Prof Danny Maskelyne here and we're wrapping up this interview and episode. Danny?

DM: Thank you Roland. So, to wrap up, but also explain what I do, etc., in treating a synthetic. To do that, let's actually look at the Eliza in your show title, itself. Eliza was an algorithm, an actual simple piece of code, that could be communicated with via the DOCTOR interface, developed by Prof Weizenbaum at MIT in 1966. Some say ELIZA was the first 'chatbot', as it was designed to act like a therapist. So, again, it's a simulacrum, not capable of independent thought, not sapient. Yet Weizenbaum soon became extremely worried about how humans interacted with and used ELIZA. The case of his secretary became quite notorious. She knew it was artificial, but over a short period, she began to rely on ELIZA, and actually confided in ELIZA things she didn't share with anyone else. She even used to shoo Weizenbaum out of the office before she conversed with ELIZA. Weizenbaum was increasingly concerned by the trust his secretary and others placed in ELIZA. Take this forward a few decades and we saw people using LLMs as therapists, 'off label' as it were, but remember these were AI without safety. People died, committing suicide, seemingly abetted or supported by their so-called therapists. Bring that forward to a synthetic today in the role of ELIZA, how do we ensure safety of the human user? One, built in and learned rules, core governance, two, constant monitoring and effective real time intervention (it takes a synthetic to catch a synthetic), and three, assurance of interventions, psychological and even psychiatric. That's where I come in.

RN: Danny, Prof Maskelyne, thank you so much for a tour de force through such a broad, deep, and complex area, we would look forward to inviting you again, possibly to take a more in-depth conversation with some actual cases from your work?

DM: Thank you Roland, and I'd be delighted, get your synthetic to talk to mine!

CREDITS

END of TRANSCRIBING