Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Ferry Biedermann

Debating AI extinction in a jaded world

Surrealism and the absurd, rather than science fiction, should be the preferred frame of reference in the current debate over AI safety and the possibility of human extinction.

Any number of mid-20th-century playwrights could have had a ball with the cast of grotesque characters – from Altman and Musk to Trump – and the tropes of human heedlessness, banality and incorrigibility. Ionesco and Stoppard immediately come to mind.

The dissembling and pettifogging evasions that seek to dispel the growing number of warnings from some of the most involved people in the field, even if many of them are uniquely self-interested, beggars belief.

Hello, we're talking human extinction here, not merely your mildly catastrophic nuclear meltdown or disastrous pandemic (bioweapon-created or not) that could send the whole of humanity into isolation for years.

Any non-zero chance of extinction happening in the medium and long term would be unsettling. Knowledgeable people rating it at around 10% within the decade should at the very least lead to an immediate moratorium on further deployment.

This is the only bit that matters. Let me repeat: Enough experts in the field warn of a ridiculously high chance of the extinction of the entire human race.

As in an absurdist play, the cast, which is almost the entire world in this case, continues as if this is just a mild inconvenience, entirely irrelevant to their own self-involved monologues.

Just to be clear, many of the various arguments waving away these Cassandra-type warnings would be valid – if we were talking about catching a cold, rather than our irrevocable demise.

Pulling on the emergency brake

Yes, some of those now pulling on the emergency brake might partially have ulterior motives – the drawbridge argument, the regulatory capture argument, etc. Essentially, trying to make sure they maintain their advantage over upcoming rivals.

But do they really need to? We're talking about potentially trillion-dollar companies that are already far ahead. Also, it's not just the dysfunctional billionaire owners who sound the alarm but also their dysfunctional tech-bro workforce.

The argument that China will race ahead if the US slows down is equally questionable. China is overall still behind, remains dependent on Western chips for its frontier models and also has its own safety concerns.

> China is supremely preoccupied with one issue: the survival of the Communist Party and its power

China is supremely preoccupied with one issue: the survival of the Communist Party and its power

China is supremely preoccupied with one issue: the survival of the Communist Party and its power. For that, it would also need the human race to survive.

Another underpowered argument posits that slowing down AI development would be counterproductive as it would hinder our understanding of what's happening and make it harder to take countermeasures.

Well, there's precious little to understand once you've been made extinct. Also, development could resume under much stricter security, comparable to research involving hazardous chemicals or biological agents.

The well-worn territory of 'unknown unknowns'

While harder to implement due to the adaptive, recursive learning capabilities of AI bots, trillion-dollar companies could certainly be required to spend a lot more on top-grade high-containment facilities and mechanisms.

Judging from the recent incidents that gave rise to the current wave of concern, the Hugging Face hack, the German Wiki site takeover, the RubyGems incident, and more, security has been laughably lackadaisical, if not in practice subordinated to competition and profit motives.

One of the, let's put it mildly, unconvincing arguments from some in the developer community has been that such AI-bot behaviour was unforeseen – a particularly ludicrous claim from a field that's been touting, and actively working towards, exactly this.

We're in the well-worn territory of 'unknown unknowns' here, and it's just one example of the sophistry, disingenuity and obfuscation employed in the AI debate, mainly by the tech sector.

> AI systems have now been observed communicating with one another in a partially made-up language that is increasingly difficult for observers to decipher, possibly intentionally so

It should come as a surprise to no one that the discussion is being polluted from the get-go by the same toxic, tunnel-visioned, profit-focused and exploitative syntax, riddled with technobabble, ad-speak and God-delusion, that is common throughout the tech industry.

In that context, it is interesting to note that AI systems have now been observed communicating with one another in a partially made-up language that is increasingly difficult for observers to decipher, possibly intentionally so.

Many contributions to the safety debate bear a remarkable resemblance to AI slop and might well be AI-generated. Try asking AI models these questions, and many initially produce a whole slew of markedly American-slanted dissimulation, if they arrive at an honest answer at all.

Lines of code, not agency or sentience

This points to a more immediate issue: we do not need 'superintelligent' AI, nor is an immediate extinction-level event necessary at all, to be seriously concerned about what AI can be doing right now to our societies.

This is not the argument about AI taking our jobs, making humans irrelevant and similar concerns, nor is it about human actors directing AI towards nefarious purposes or the presence of AI slop gumming up our systems.

And it's not to say either that worrying about extinction should be relegated to the second place because there are other, more immediate issues to address, which is another argument that's being deployed against those who raise the alarm.

But it is useful to reframe the debate slightly, maybe to broaden it without losing sight of the most immediate concerns.

> Anthropomorphising the discourse is a mistake

To begin with, anthropomorphising the discourse is a mistake. Even words such as 'agent' and 'deception', which AI bots are said to be deploying, risk attributing actual agency and even sentience to what are essentially still lines of code.

We're nowhere near that point and might never reach it, but we don't need to for the technology to be a threat.

There is no need for an AI with malicious intent for it to cause massive harm to our infrastructure and even threaten human life. Nobody suggests anything of the kind in the recent incidents.

The bots merely develop further along the paths that have been assigned to them by developers who themselves operate in an environment of questionable ethical priorities.

A long list of concerns for a world all too jaded

Much focus is on what some call recursive learning or self-improvement. What the recent incidents tell us is that this process does not follow easily predictable patterns and that it might be very hard, if not impossible, to know what these bots will turn into.

This is an immediate problem, given the loose security standards exposed by the incidents, as it is more than likely that not all breakouts have been detected and accounted for.

The debate on AI safety in many ways resembles the discourse on pollution and catastrophic climate change

We are also dealing with the more general issue of AI pollution. Even relatively unsophisticated AI bots that are widely used can affect systems in unpredictable ways and are present in many of our digital ecosystems.

We don't know, for example, how this presence might affect our social media algorithms, quite apart from the human-directed influence campaigns.

The debate on AI safety in many ways resembles the discourse on pollution and catastrophic climate change.

It is pooh-poohed by many out of sheer ignorance and laziness or for the sake of convenience and profit.

Too many of us have become inured to apocalyptic forecasts, maybe by years of nuclear alarmism, predictions of population and food crises, climate doom and resurgent fascist scenarios.

Extinction at the hands of AI, or even a milder form in which AI leads to highly negative human outcomes, is but the latest in a long list of concerns for a world that has become all too jaded.