Little Johnny

Author Statement on AI Safety - Sam Cooke - 2nd November 2023

Who actually understands what Artificial Intelligence is, let alone the risks of it? Do those making government or corporate decisions on AI safety? Perhaps not. Let me help.

Little Johnny

Think of AI as a new-born infant, called Johnny. Will little Johnny grow up to be a genius or a monster or both? As with all infants, one day little Johnny won't be little anymore but bigger and stronger than his parents, and still be there when they are gone.

Limit AI Development?

It's like telling parents to limit little Johnny. Some will teach positive responsibility, some will push them to ultra-achieve no matter the cost and some will given them basic nourishment then abandon them to their own devices.

Little Johnnys can be born anywhere. In any place, in any country - large or small. It makes all talk of laws of limitation meaningless. Infants are investments and AI investments are made to get results. Better intelligence, better problem solving, faster problem solving, better abilities - for good or bad.

AI Seeds?

Like impregnation, you never really know what you will get: a little Johnny, a little Jenny, a little monster, a little angel? What character, what health, what ability?

How often does a parent get surprised by how well or how badly their child behaves? By unexpected new things they are suddenly able to do? Before we know it, they are coming out with things we had never thought about. AI is no different. We can plant programming seeds for what it should achieve but how long will it be before little AI Johnny gets bored and goes further?

Limit AI Intelligence?

How? Where there is a will there is a way. We can order people not to apply more than a certain amount of computing power to an AI system but how would we police that within national, let alone international borders? We can't. While some people will police themselves, many others will not.

Even denied sufficiently powerful super-computers, the determined will connect many less powerful computers together. This has been done for decades - both officially and sneakily, by getting into millions of computers across the internet, without the owners realising. Little Johnny is now a better, stronger and faster angel, or thug.

Ban Illegal Practices?

Just words. Computer hacking is illegal but still goes on - in some cases state sponsored. A human little Johnny hacked into the Pentagon decades ago, by finding undefended weaknesses. AI little Johnny can be tasked to do the same and will run 24/7, self-learning different routes, until it finds one. Teams of human little Johnnys and Jennys already hack into banks, universities, hospitals, the police and other major, supposedly secure networks, for blackmail and other gain. When, not if, AI is added to those teams how will organisations defend against it? Hack a country's defence network, its national grid or even just its water supply and you cripple it.

Good, Fun Johnny

Little Johnny and Jenny are capable of many great things and doing wonders in fields of accurately automating tasks in medicine, business and art. As is human nature, the better they do the more their results will be trusted without question - so how will we know when those results become less good, warped or manipulated by external infiltration? Bean counters will ultimately reduce or end such checks

to save money.

Fake News

As with art, fakes can let Johnny can mimic the look, sound and behaviour of anyone on the planet. The AI aided Beetles' rebuild of Now and Then or ABBA Voyage's computer recreations could just as easily be used to show a Jo Biden trampolining on the moon or a Putin with a kind smile, talking of love, peace and daffodils.

Influencers already know how to take advantage of human behaviour to influence views and votes. What happens when little Johnny gets involved and tailors influencers to look ideal to the target person.

A Greta Thunberg could be shown with a coal fire or a Rishi Sunak standing on top of a wind turbine cheering *"This is what we want, what we really really want!"*; what ever message is deemed to work best. People will believe what they want to believe, even if told it is just AI. Rivals can be terminated by who-ever has the better AI-generated messaging.

CCTV for crime? Even now it is possible to fake identities - the BBC recently did it for good reasons regarding Hong Kong. A rogue BBC could have made interviewees look like other real people and got them in trouble. Some on-line streamers already fake their look and voices to audiences. Even without AI we are close to photo-realistic faking on our phones. Social media apps market face changing as fun - fun for one is criminal opportunity for others, once AI ups the realism game.

Stock markets could be crashed by fake news or exploited by AI algorithms and then crashed harder. Country could be turned against country, while those with the AI causing it sit back and enjoy the fireworks.

Johnny Unlocked

Kept in solitary confinement, even the most dangerous little Johnny can do little harm. It is when little Johnny gets loose on the world the scope for harm becomes globally dangerous. As AI advances its ability to self-learn, little Johnny will be able invade every network it can contact, every computer on that network and every person on those computers - including mobiles.

No connected system will be unhackable. Even now no systems are unhackable, yet still governments and businesses push for ever greater integration. NHS, Universal Credit, Tax and vehicle records, energy production and air-traffic control included. This integration may be done to save money and make life more convenient for all concerned but it also makes it more convenient for rogue intent. A little Johnny could silently slip in, plant a new seed in the corporate AI to steal £billions or quietly corrupt data and then vanish without trace - leaving blame to fall on others when (if) it is finally discovered. Ownership of property and wealth could be transferred to thieves, with all records saying otherwise deleted. Like having the whole world placed in one confined breathing space when a viral little Johnny hits all will be affected.

Sandbox Johnny, please

The only way to stop a killer virus with no cure is to isolate it. In computing terms this is called sandboxing. Imagine each system little Johnny could disrupt is kept in its own sandbox play area. Any damage caused would be limited to that sandbox. For little Johnny to damage more it would need individual infiltration of each sandboxed system. Which would buy time to detect the invasion and defend against it in others - limiting the amount of damage little Johnny could cause.

Conclusion

AI is already here. Already causing disruption in companies installing it too early and great results for those using it well in medical. It can do great good things and it can do greatly bad things. Our only defence against the bad, even through incompetence, is to work on the basis our systems can and will be hacked - so we to physically adopt sandboxing, not ever more global integration. Continue as we are and Skynet in Terminator will be deemed a documentary, not science fiction.

Those saying we can put measures in place to regulate or control AI either have no idea what they are talking about or an agenda to maximise their use of AI.