Australian Government
Chief Scientist

DR ALAN FINKEL AO

Creative Innovation Summit 2019

Keynote address

Laws of Robotics… for humans

Tuesday 2nd April 2019

Sofitel
MELBOURNE

It feels appropriate to begin today's proceedings with a philosophical question.

Can machines *create*?

Or, more precisely, can artificial intelligence, AI, be creative?

Sit with that question for a while – as we turn to the real substance of our discussions today.

And that's not merely the *concept* of creative machines, but the question that ought to follow: are *we* creative enough to respond?

It's a timely question, for all sorts of reasons... but for me, in 2019, it's got a special resonance.

That's because we're about to pass a very significant milestone: the hundredth anniversary of the birth of the science fiction writer Isaac Asimov.

He was born in a little village on the Western fringes of Communist Russia.

His family emigrated to the United States when Isaac was three.

They settled in New York, and they went into business – running a candy shop.

Remember this detail, because it turns out to be important.

Asimov graduated from high school at age 15, applied for medical school but failed, and ended up in chemistry.

So far, so ordinary.

And then, in 1941, when Asimov was just twenty-one, he did something truly remarkable.

He became the first person to use the word "robotics" in print, in a story he submitted to a magazine.

And we *know* he was the first, because it was officially confirmed by the US Congress in 2002, when it established National Robotics Week in Isaac Asimov's honour.

And in the following year, 1942, Asimov did something even more remarkable: he came up with the iconic Three Laws of Robotics.

- A robot must not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey orders given to it by human beings.
- A robot must protect its own existence.

In the event of a conflict, the first rule trumps the second, and the second rule trumps the third.

It might seem extraordinary that a chemistry student, just 22 years old, could pen something that stands to this day as technology's version of the Ten Commandments.

But Asimov would always say that he hadn't done anything particularly extraordinary at all.

He'd just spent a lot of time in... yes... his parents' candy shop.

Which *also* sold science fiction magazines.

Which the young Isaac devoured in the way that other children gulp down chocolate.

By age 22, he had decided that all the robots in fiction were boring.

It was same old thing, every time: Humans build robots. Then everybody dies.

And Asimov, he was over it.

He *hated* to be told that creativity was bad and humans shouldn't meddle.

But more fundamentally, he thought the robot apocalypse plotline didn't actually tell us anything useful about our relationship with technology at all.

Yes, a technology like a robot *could* be viewed as inherently dangerous and largely unpredictable – but *lots* of things are inherently dangerous and unpredictable, like, for example, other people.

And people, for the most part, got along, because every functioning society teaches its children a human version of the three Robot Laws:

- Don't hurt other human beings.
- Obey lawful instructions.
- And don't hurt yourself.

So Asimov thought he wasn't so much inventing robot laws as codifying, literally, the basic principles that all humans hold in common, so the robots could be integrated into our society.

He *also* saw that, with those same three principles, we had made useful tools of any number of potentially dangerous things.

Why would humans – the same species who had tamed fire, and cars, and electricity, and aeroplanes, and medicines, and explosives – suddenly forget everything they'd learned when it came to robots?

They wouldn't – it was completely implausible.

Now a lesser brain than Isaac Asimov's might have given up and put down the pen right there, on the basis that if there's not going to be a robot apocalypse then where's the story?

But for Asimov this is where the real interest began – *not* with an apocalypse, but with human beings actually having to *live* with their creations. Working out the details, with limited information, and conflicting needs and priorities, every day. And both succeeding brilliantly and stuffing it up!

The Three Laws were just the start. What followed – how humans decided to interact with their robots, and why – was always the substance. The story.

And that's why I find myself turning back to Asimov, as I look around the world and see that yes... that's exactly the story we're living today.

Tech developer Marc Andreesen, who created Mosaic, the web browser that popularised the world wide web, said in 2011 that "software is eating the world".

By now, it's licking the plate.

And in many cases, the outcome isn't just good – it's great!

To give you just one example, consider the use of AI for IVF.

Typically, clinics wait for the newly fertilised eggs to develop over four or five days into embryos, before the doctors decide which of the batch to implant.

They judge which embryo gives the best shot at a successful pregnancy by their appearance.

But human doctors can't watch an embryo constantly – 24/7 – for five days straight.

AI can.

And human doctors can't be trained on thousands and thousands of hours of time-lapse footage of embryo development.

AI can.

And then the AI can help the doctors to make better decisions.

Well, that AI exists. For the record, her name is Ivy – and she's Australian, born and bred in Sydney.

Think about it. Really think about it. We're raising a generation of humans who won't just grow up with AI from birth – they're growing up with AI from conception.

In so many places, science fiction is coming to life.

So, turn the page – exercise that human creativity to think ahead – and what happens next?

We could sleepwalk into a society of mass surveillance – an Orwellian dystopia. For me personally, that's the scariest scenario.

At the other end of the spectrum, we could scare ourselves into kneejerk technology abstinence – and see how long we can live with an Amish level of self-denial.

Or we could realise the hope of this summit, and find the Goldilocks option in the middle: a society which integrates its robots well.

At last year's summit I asked you to think about what it would mean for humans and AI to play nice and get along.

This year, in honour of Isaac Asimov's birthday, I thought I'd come back to you with my thoughts on how we ought to approach that interaction – in the form of four new Laws of Robotics.

But there's a difference with these laws: they don't apply to the robots – they are directed at the *humans*.

And, just because 2019 is *also* the 150th birthday of the periodic table, I've named them after elements.

Starting with the classic: The Golden Rule.

This rule is borrowed from the Secretary of the Department of Home Affairs, Mike Pezzullo, whose office deals with the practicalities of AI integration every day.

It's this:

No robot or artificial intelligence system should ever take away someone's right, privilege or entitlement in a way that can't ultimately be linked back to an accountable human decision-maker.

It's elegant, and compelling: a clear statement of what human justice in 2019 demands.

By following this rule, we will not detach human loss and human suffering from human agency, judgment, and empathy.

The second rule, which I call the Carbon Rule, given that human beings are carbon-based life-forms.

And the rule is straightforward: if machine intelligence is advancing, then human intelligence ought to do the same.

There's a belief in some quarters than AI means outsourcing the work of our brains, so as the robots smarten up, the humans can dumb down.

I fundamentally disagree.

We need *more* education in order to integrate robots, not less.

We need *more* discussion of philosophy and ethics, not less.

We need *more* creativity, in politics and in literature and in schools and in the media, not less.

Asimov said it himself: *"While knowledge can create problems, it is not through ignorance that we can solve them".*

So the Carbon Rule: don't dumb down, skill up.

The third rule, which I call the Argon Rule.

Argon: a noble gas, meaning that it's basically unreactive.

It's not like fluorine.

Fluorine blows up on contact with water because it’s always so hungry for electrons.

Fluorine is extremely grabby.

Not Argon!

So the Argon Rule is a plea to technology developers and adopters: *don’t be greedy*.

In your reach for data or profits, don’t run so far ahead of the community’s interests that you lose the privilege of a long and generous leash.

And finally, the Platinum Rule.

Every machine should have an off-switch. And an off-switch is useless unless there are humans who are trained to know when to use it.

Gold, Carbon, Argon, Platinum.

Four Human Laws of Robotics, to sit alongside the classic three.

And that just leaves me with time for one final exhortation to everyone taking up the challenge of Creative Innovation today – enjoy the summit and...

May the Force be with you.