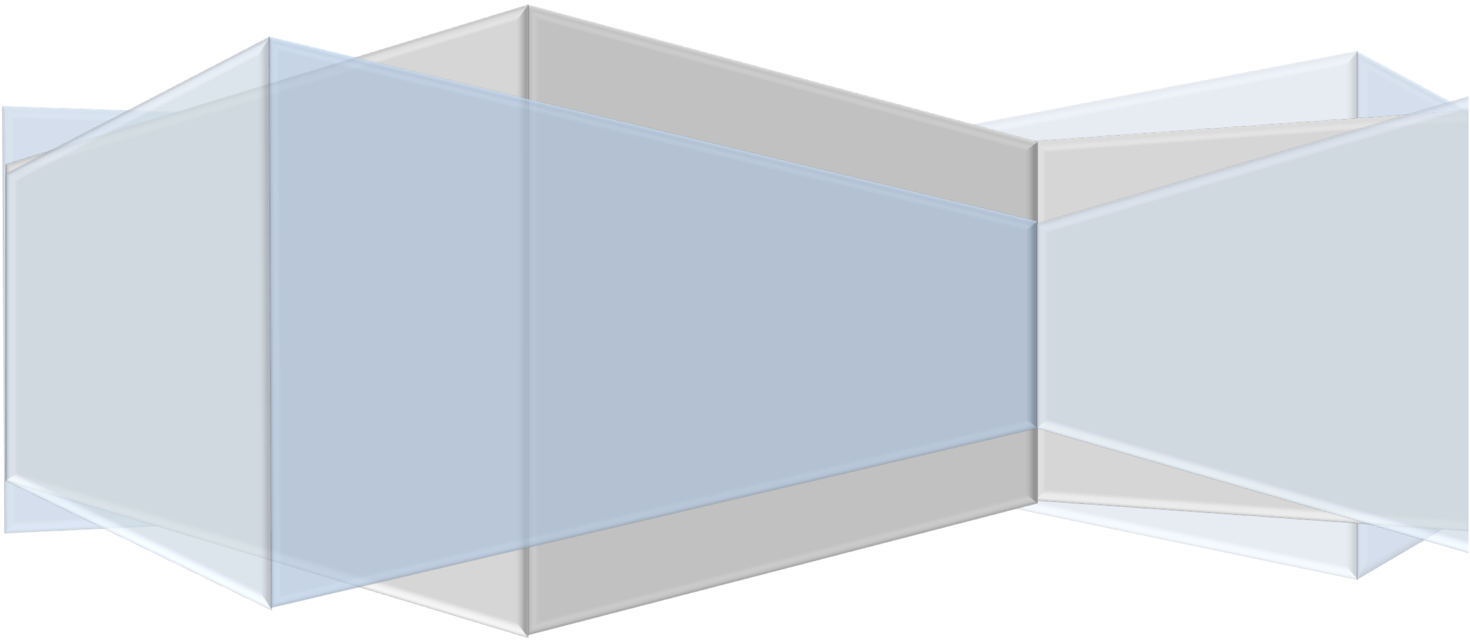
Rustat Conferences ◦ Jesus College ◦ Cambridge

Machine Superintelligence & Humanity

Jesus College,
Cambridge
Thursday, 2 June 2016

Rapporteur: Nathan Brooker



Rustat Conferences

Machine Superintelligence & Humanity

Jesus College, Cambridge Thursday, 2 June 2016

Contents

Conference Report:

Rustat Conferences
Jesus College
Cambridge

The Rustat Conferences are an initiative of Jesus College, Cambridge, and chaired by Professor Ian White FREng, Master of Jesus College. The conferences provide an opportunity for decision-makers from the frontlines of politics, the civil service, business, the professions, the media, and education to exchange views on vital issues of the day with leading academics.

Founded in 2009, Rustat Conferences have covered a variety of themes including: *The Economic Crisis; The Future of Democracy; Cyber Security; Manufacturing in the UK; The Geopolitics of Oil and Energy; Drugs Policy; Organisational Change in the Economic Crisis; Cyber Finance; The Understanding and Misunderstanding of Risk; Food Security; Transport and Energy, Inequality, Big Data,* and *the UK North South Divide.* In addition to acting as a forum for the exchange of views on a range of major concerns, the conferences provide outreach to a wider professional, academic, student and alumni audience through the publication of reports. The conferences are named after Tobias Rustat (d.1694), a benefactor of Jesus College and the University.

Chatham House Rule and Rustat Conference Report

Please note the conference is conducted under the Chatham House Rule. The conference report will however reveal the identity and affiliation of speakers and discussants unless they request otherwise. The procedure is as follows: the Rustat Conference rapporteur, Nathan Brooker, will circulate a draft version of the report and anyone mentioned or quoted in it may remove the attribution. Once this procedure is complete the report is published via the Rustat Conferences website.

#Superintelligence #Rustat @JesusCollegeCam

Rustat Conferences

Rustat Conferences
Foundation Members

The Rustat Conferences are supported through a mix of sponsorship and a membership scheme that was launched in 2013-14 - details of this can be found at www.Rustat.org. We are very grateful to the Rustat Conferences Foundation Members for their generous support:

Dr James Dodd - James's career has concentrated on the financing and management of companies in the fields of telecommunications and technology. He serves on a number of boards in these areas and is active in supporting a number of academic projects and charities.

Harvey Nash is an executive recruitment and outsourcing group. Listed on the London Stock Exchange, and with offices across the world, we help organisations recruit, source and manage the highly skilled talent they need to succeed in an increasingly competitive and innovation driven world.

KPMG is a global network of professional firms providing Audit, Tax and Advisory services. It has more than 155,000 outstanding professionals working together to deliver value in 155 countries worldwide.

Laing O'Rourke is a privately owned, international engineering enterprise with world-class capabilities spanning the entire client value chain. We operate an integrated business model comprising the full range of engineering, construction and asset management services delivering single-source solutions for some of the world's most prestigious public and private organisations.

McLaren Racing Ltd has a reputation for efficiency and professionalism. Working within a fast-paced environment and to the highest standards, our highly skilled workforce operates primarily in the areas of manufacturing, engineering and race team as well as logistics and support.

Mr Andreas Naumann is a senior executive in the financial industry. Outside the professional sphere, he is keenly interested in subjects like urbanisation, youth unemployment, education and foreign policy. He supports the Rustat Conferences as a private individual.

Sandaire Investment Office - SandAire and Lord North Street came together in April 2014 to combine their businesses, both of which specialise in looking after the investment assets of very wealthy families, charities and endowments.

Maria and David Willetts

RUSTAT CONFERENCES
JESUS COLLEGE, CAMBRIDGE

Machine Superintelligence & Humanity

Thursday, 2 June 2016

8.45-9.30: Registration and Refreshments: Prioress's Room, Cloister Court, Jesus College. Move to Upper Hall for Conference

9.30-10.15: Professor Ian White *Master, Jesus College; Van Eck Professor of Engineering, University of Cambridge; Chair, Rustat Conferences*

INTRODUCTION
Chair: Professor Lord Martin Rees *Co-Founder, Centre for the Study of Existential Risk; former Master, Trinity College, Cambridge; former President, Royal Society*

***From Artificial Intelligence to Superintelligence**:* Dr Hermann Hauser *Co-Founder, Amadeus Capital Partners; Honorary Fellow, King's College, Cambridge*

***AGI and Existential Risk**:* Professor John Naughton *Senior Research Fellow, CRASSH, University of Cambridge; Emeritus Professor of the Public Understanding of Technology, Open University*

10.15-11.30: SESSION 1: *Creativity and Imagination in Humans and Machines*

Chair: Professor Margaret Boden *Research Professor of Cognitive Science, University of Sussex*
Dr Demis Hassabis *Co-Founder and CEO, DeepMind (Google)*
Professor Simon Colton *Professor of Computational Creativity, Goldsmiths, University of London*

11.30-11.50: COFFEE: Gallery, Upper Hall

11.50-13.15: SESSION 2: *Human-AI Equivalence and Beyond: Personhood, Consciousness, Values*
Chair: Professor Murray Shanahan *Professor of Cognitive Robotics, Imperial College London*
Professor Tim Crane *Knightbridge Professor of Philosophy, University of Cambridge*
Professor Sarah Coakley *Norris-Hulse Professor of Divinity, University of Cambridge*

13.15-14.15: LUNCH: Master's Lodge and Prioress's Room

14.15-15.30: SESSION 3: *Ethics, Regulation and Governance*
Chair: Professor Huw Price *Academic Director, Leverhulme Centre for the Future of Intelligence Bertrand Russell Professor of Philosophy, University of Cambridge, Co-Founder, Centre for the Study of Existential Risk*
Dr Jatinder Singh *Computer Laboratory, University of Cambridge*
Professor Murray Shanahan *Professor of Cognitive Robotics, Imperial College London*

15.30-16.45: SESSION 4: *Anthropology and Theology in An Era of AGI*

Chair: Professor Andrew Briggs *Professor of Nanomaterials, University of Oxford*
Dr Andrew Davison *Starbridge Lecturer in Theology and Natural Sciences, University of Cambridge*
Dr Timothy Jenkins *Anthropologist, Fellow, Jesus College, Cambridge*

CONCLUDING REMARKS: The Master, Jesus College**,** Professor Ian White

16.45-17.15 TEA: Gallery, Upper Hall

RUSTAT CONFERENCES
JESUS COLLEGE, CAMBRIDGE

Machine Superintelligence & Humanity

Thursday, 2 June 2016

Background

Developments in the fields of Artificial Intelligence have long given notice that the products of technology have the scope to outstrip some or even many of the functions performed by human intelligence. There has been talk of a “tipping point”, where, as Professor Huw Price puts it, “intelligence escapes its biological constraints”, and beyond which the consequences may well be both unfavourable and irreversible.

There are wide-ranging arguments about the pace and extent of these developments; also about the advantages and drawbacks across a broad span of fields involving health, economics, leisure, employment, education, security, defence. The new Leverhulme Centre for the Future of Intelligence, based in Cambridge with partners at Imperial, Oxford and Berkeley, aims, as they put it, to 'to bring together some of the best of human intelligence, to ensure that we humans make the best of machine intelligence'.

The Rustat Conference on Superintelligence and Humanity aims to explore a “qualitative understanding” of future developments in Artificial Intelligence alongside the more pragmatic considerations of “existential risk”. By encouraging the arts and the humanities to join the discussion, we hope to offer additional dimensions to the current debates and critiques, at the same time suggesting directions for associated research, further meetings, and publications.

The conference brings together researchers in fields of AI, robotics and computer science with scholars working in the fields of philosophy of mind, history of ideas, political science, anthropology, psychology, literary studies, sociology of religion, and theology. The future of Artificial Intelligence raises questions about the place of humans in the world beyond the “tipping point”. Inevitably this prompts reflections on the implications for the moral, cultural, and spiritual dimensions of human life.

Acknowledgements

This meeting is organised in conjunction with the Centre for Study of Existential Risk and the Leverhulme Centre for the Future of Intelligence. We are grateful for the support of the Hauser-Raspe Foundation, and the Rustat Conferences Foundation Members.

#Superintelligence #Rustat @JesusCollegeCam

Participants at Rustat Conference
on Machine Superintelligence & Humanity
Jesus College, Cambridge - Thursday, 2 June 2016

- Dr Denis Alexander, Emeritus Director, Faraday Institute, Cambridge
- Professor Jean Bacon: Distributed Systems, Computer Laboratory Jesus College, Cambridge
- Professor Jeremy Baumberg: Nanoscience, Physics, Jesus College, Cambridge
- Kristin Bayne: Deputy Development Director, Jesus College, Cambridge
- Professor Margaret Boden: Informatics, Cognitive Science (Sussex)
- Professor Pete de Bolla: English Literature (King's College, Cambridge)
- Professor Andrew Briggs: Nanomaterials (Oxford) TWCF
- Dr Tom Chatfield: author, games designer
- Dr Lucy Cheke: Neuroscience (Cambridge)
- Professor Robert Cipolla: Robotics, Engineering (Jesus College, Cambridge)
- Stephen Clark: Reader, Natural Language Processing (Cambridge)
- Professor Sarah Coakley: Philosophy of Religion (Murray Edwards, Cambridge)
- Professor Simon Colton: Computational Creativity Group (Goldsmiths and Falmouth)
- Professor Alastair Compston: Neuroscience (Jesus College, Cambridge)
- Professor Ann Copestake: Computational Linguistics, Cambridge
- George de Courcy-Wheeler: Deputy Chief Investment Officer, Sandaire; Rustat Conferences Foundation Member
- Professor Tim Crane: Philosophy (Peterhouse, Cambridge)
- Dr Andrew Davison: Science and Religion, Starbridge Lecturer (Corpus Christi, Cambridge)
- Dr Sara Dillon: Faculty of English (Cambridge)
- Dr James Dodd: Financier and Technologist, Rustat Conferences Foundation Member
- Revd Dr Paul Dominiak: Dean of Chapel, Jesus College, Cambridge
- Albert Ellis: CEO, Harvey Nash; Rustat Conferences Foundation Member
- Dr Fiona Gatty: Nanomaterials (Oxford) TWCF
- Dr Sean Gourley: Founder and CTO, Quid
- Dr Marta Halina: History and Philosophy of Science, Cambridge
- Dr Demis Hassabis: CEO and Co-Founder, DeepMind (Google)
- Helen Harris: Communications Manager, Jesus College, Cambridge
- Dr Hermann Hauser: Co-Founder , Amadeus Capital Partners
- Professor Stephen Heath: English Literature and French Culture (Jesus College, Cambridge)
- Marc Hendriks: Chief Investment Officer, Sandaire; Rustat Conferences Foundation Member
- Dr Julian Huppert: POLIS (Cambridge)
- Dr Katherine Jenkins: Philosophy (Jesus College, Cambridge)
- Dr Tim Jenkins: Anthropology of Religion (Jesus College, Cambridge)
- Dr Peter Jordan: Nanomaterials (Oxford) TWCF
- Dr Duncan Kelly: POLIS (Jesus College, Cambridge)
- Professor Tim Lewens: History & Philosophy of Science (Cambridge)
- Richard Muirhead: General Partner, Open Ocean; Co-Founder and Chairman, Firestartr
- Professor John Naughton: Public understanding of Technology; CRASSH (Wolfson College, Cambridge)
- Jonathan Neale: CEO, McLaren; Rustat Conferences Foundation Member
- Sumit Paul-Choudhury: Editor, New Scientist
- Dr Julia Powles: Postdoctoral Researcher, Faculty of Law and Computer Laboratory, Cambridge

- Professor Huw Price: Academic Director, Leverhulme Centre for the Future of Intelligence, Bertrand Russell Professor of Philosophy, Co-Founder Centre for the Study of Existential Risk (Cambridge)
- Nick Ray: Architecture and intellectual history (Jesus College, Cambridge
- Professor Lord Martin Rees: Cosmology, Co-Founder, Centre for the Study of Existential Risk; former President, Royal Society (Trinity College, Cambridge)
- Richard Sandford: Government Office for Science, Head of Horizon Scanning
- Dr Simone Schnall: Experimental Psychology (Cambridge)
- Dr Ulrich Schneider: Department of Physics (Cambridge)
- Dr Abigail Sellen: Microsoft Research, Cambridge
- Dr Andrew Serazin: President, Templeton World Charity Foundation (TWCF)
- Professor Murray Shanahan: Artificial Intelligence and Robotics (Imperial)
- Dr Jat Singh: Ethics, Regulation, Cambridge Computer Laboratory
- Dr Beth Singler: Faraday Research Institute (Cambridge)
- Dr Francis Spufford: Author, Creative Writing Programme, Goldsmiths
- Dr Jaan Tallinn: Co-Founder, Skype; Co-founder, Centre for the Study of Existential Risk
- Roger Wagner: Artist
- Professor Marc Warner: Quantum Physics and AI (Physics, Harvard)
- Paul Westbury: Technical Director, Laing O'Rourke; Rustat Conferences Foundation Member
- Adrian Weller: Machine Learning, CSER, CSAP, University of Cambridge
- Professor Ian White: Photonics, Engineering; Master, Jesus College, Cambridge; Chair, Rustat Conferences
- Margaret White: Jesus College, Cambridge
- Professor Tim Wilkinson: Engineering (Jesus College, Cambridge)
- Professor Peter Williamson: International Management (Cambridge Judge Business School) ; Advisory Board, Rustat Conferences

Rustat Conferences, Jesus College, Cambridge:

- Jonathan S. Cornwell: Director, Rustat Conferences info@rustat.org
- Rustat Conferences Rapporteur: Nathan Brooker (Jes Coll 2009), Financial Times editorial ncbrooker@yahoo.co.uk
- Rustat Conferences: Dr Tudor Jenkins, (formerly AI, Sussex University and Ecole Normale, Paris)
- Jordan Burgess: MPhil in Machine Learning, Speech and Language Technology (Cambridge)
- Mark Mawdsley: his PhD applies machine learning to a multi-objective optimization to improve the core design for the world's most abundant nuclear reactor (Cambridge)
- Rustat Conferences Founder: John Cornwell, Director Science & Human Dimension Project (Cambridge) jc224@cam.ac.uk

RUSTAT CONFERENCES
JESUS COLLEGE, CAMBRIDGE

Machine Superintelligence & Humanity

Thursday, 2 June 2016

Executive Summary

Over the course of the next century, the development of superintelligent AI systems will pose great challenges to humanity. It will ask questions of us like no other technological advancement has done in history – questions about what it means to be human; how our society functions and how it ought to. It will ask us questions about how we should treat AI systems and how we would wish to be treated by them.

While the arrival of truly superintelligent machines is still downstream of us, we should ask ourselves these questions now. Such is the rapidity of technological improvement in this field, we may only get to ask them once.

The day's discussion largely covered three areas:

1. Capacity

Calculators have had superhuman abilities for 40 years, so what is meant by the term 'superintelligent' machines? For many, it comes down to a question of range:

> *"A superintelligence an intellect that is much smarter than the best human brains in practically every field including scientific creativity, general wisdom and social skills".*
>
> *Nick Bostrom (2006)*

AI systems have been developed to beat humans at chess, Jeopardy and "Go"; but these systems are not necessarily intelligent in the way a human mind is. The developers of 'Watson', the AI system that beat human competitors at Jeopardy, claim its learning process mimics our own. But does its capacity to read millions of unstructured documents in seconds preclude it from being truly like us? Humans also display various types of intelligence – could an AI machine ever be emotionally intelligent? Other areas where artificial and human intelligence may diverge include:

- **Personhood**: discussions of AI rarely take into account the granularity of their individuation. When we use the term AI, are we referring to a human-like entity or a giant, disembodied superintelligence? This is crucial to any discussion on its sense of self (if any) and how its sensors would perceive the world.
- **Consciousness**: while machines are becoming more and more intelligent, they are not getting any closer to having consciousness. Furthermore, the capacity for consciousness may be an attribute we are able to build into software. After all, we can conceive both of a human-like entity with little or no consciousness or interior life, and a non-human like 'hive-mind' AI that with a capacity for consciousness that is far greater than our own.

- **Creativity/intuition**: projects such as AlphaGo have exposed the limitations of brute force tree searches. In response, more intuitive, value-aided judgements have been developed. These could be the first steps towards creating a genuinely creative AI system. That journey will be complete when software's creativity is taken so seriously that it can itself enter the debate over what 'creativity' means.

2. Risk

The risks posed by AI have had currency for 50 years. But the immediate threats have largely been overstated. Areas discussed include:

- **Liability**: knowing who is legally responsible is difficult to determine in the algorithmic world – system transparency and the ability to implement constraints are vital to weed out biases and error. When machines are learning from other machines, working out who is liable is even more complicated.
- **Labour**: within 100 years machines will potentially threaten more than 50 per cent of jobs in the labour market – moving from blue collar jobs (truck driving) to white collar jobs (legal clerks). The unprecedented pace of AI development means technological advancement in this instance may not throw up new, replacement jobs as it has done during previous periods of upheaval, such as the Industrial Revolution. As jobs are lost to AI advancement – and they certainly will – we must decide what kind of a society we want to live in. Should we encourage a basic state income? Or will this promote despondency? An expanding social sector, with more carer and custodian roles, may provide people with the kind of dignified employment that is beneficial to them and to society.
- **Existential**: predictions about the future of AI are often alarmist – even apocalyptic – in their moral implication, but they are often based on too-narrow an understanding of ethics. Often they do not take into account the following:

  (a) Ethics is not strictly utilitarian or consequentialist
  (b) Ethics entails virtue, or 'learning to be good'
  (c) Computers are incapable of being 'evil'

  Computers, like children, need to be taught a value-system, but values are regional and not universal. Even if we could solve the “value loading” problem of AI, we would still be confronted with the problem of which values to load. Therefore, more emphasis should be placed on the ethical lives of the machine makers.

3. Implication

Humanity will be affected by the development of superintelligent AI systems, but we cannot be supplanted from our position in society. Nor will our 'humanness' ever be diminished. Even where machine learning gives computers the ability to create their own software, at root AI is a human creation and therefore something of humanity will remain contained within it. We should therefore concern ourselves more with the ways AI will be like us, rather than where it will be different to us. If we are to deal with such technological advancements successfully, we should look to the humanities and the social sciences. These areas have experience in classifying and contextualising what is new. They will be the best tools we have to cope with a rapidly changing world.

INTRODUCTION: Machine Superintelligence & Humanity

Chair: Professor Lord Martin Rees Co-Founder, Centre for the Study of Existential Risk; former Master, Trinity College, Cambridge; former President, Royal Society

From Artificial Intelligence to Superintelligence: Dr Hermann Hauser Co-Founder, Amadeus Capital Partners; Honorary Fellow, King's College, Cambridge

AGI and Existential Risk: Professor John Naughton Senior Research Fellow, CRASSH, University of Cambridge; Emeritus Professor of the Public Understanding of Technology, Open University

Professor Lord Martin Rees

There will come a point in the future where machine learning will allow computers to develop human or superhuman capacities. To prepare for that eventuality, we should ask ourselves the following questions:

- How will these machines experience the world?
- What goals will they have? And who will set them?
- Could an artificial intelligence be ethical in any human sense? Could computers ever feel self-aware?
- What will computers do to us? And what will our obligation be to them?

Dr Herman Hauser

"From Artificial Intelligence to Superintelligence"

For a discussion about AI, a good working definition of intelligence is: "knowing what to do next". It can be deconstructed into the following parts:

- Knowing: having predictive models
- What: choice optimisation
- Do: skill set
- Next: time sequence

All to achieve a particular goal

The simplest form of intelligence that fits this model is the heliotropium. It has knowledge: it knows that the sun's position can change; and can locate its position at a given time. It has a choice – albeit a limited one – to follow the sun. It has the capacity for action: it can alter the concentration of auxin in its cells that, over the correct timeframe, causes one side of the plant stem to stiffen and one side to slacken until, ultimately, it reaches its goal and points toward the sun. In this instance, it is a goal set by evolution.

A monkey's intelligence can be similarly deconstructed. It has knowledge of and can identify food. It can identify predators and can learn their habits. When it does spot a predator it can choose from a number of responses: fight or flight, for example. It has a skill set to help it achieve this: run, climb, fight – all enacted over a certain timeframe. But while a monkey's intelligence is far greater than a plant's, its goals are still set by evolution.

While human intelligence can still be applied to the monkey model, the fundamental difference is that out goals can be self-set, rather than being dictated to us by evolution.

Self-determined goals are often thought of as the first step towards creating a superintelligence. But what do we mean by "superintelligence"?

'We have had calculators that have been superhuman in their abilities for 40 years: what do we mean by superhuman in this context?'

- Lord Martin Rees

A superintelligence can apply its thinking to more situations than the human intelligence can. It has a broader and deeper knowledge base. It can also do more things. The fundamental question, then, is will we be able to set its goals or will the superintelligence learn to set its own?

But this question is not for now. At this moment in time, superintelligences are not yet "smarter" than humans. Humans are better at spotting and exploiting similarities and differences between complex things.

AI has beaten humans at chess, Jeopardy and "Go"; it is better at speech recognition than humans, better at face and object recognition; we will soon have driverless cars. There is an assumption then that superintelligences will be better than humans in every respect within the next 100 years. But there are warnings. When that happens, machines will potentially threaten more than 50 per cent of jobs in the labour market.

Professor John Naughton

"AGI and Existential Risk"

The idea that superintelligent machines might pose an existential threat to humanity is not new. It has had currency among academics and commentators since it first appeared in a paper by I J Good in 1965.

In that time, the appetite for AI has come in cycles. Of late, we have become obsessed with it. And we have become obsessed with a very narrow definition of it: a combination of machine learning and big data.

But there are many different types of intelligences: some people are very good at calculations, for example, and some are good at empathy.

If we can create superintelligences, can we create their value systems? Or would they end up keeping us as pets?

Certain AI professionals are optimistic, believing these machines would have value-systems aligned with our own. Except human values are culture-specific and not universal.

What is more, the commercial entities championing AI have a frighteningly tin ear when it comes to any question about values. Theirs, it seems, is technocratic thinking on steroids.

Discussion

One delegate drew a comparison between the risks in developing a superintelligence and the risks in having a child. We hope our children will share the same values as us, but we don't know they won't isolate or rise up and kill us. We form children's values by bringing them into out own culture; can we do the same for machines?

The panel thought inculcating a value-system in superintelligences was important but pointed out the ability differential between the 'parent', 'child' and superintelligence: a human child has a very similar genetic set up to its parents, a superintelligence does not. In short: a man can lift a 100kg, not 100 tonnes. A superintelligence can be much more powerful than its creator.

'We have had AI winters before, but now we are in the full bloom of an AI summer'

- Professor John Naughton

In reference to Dr Hauser's talk, a delegate said that there was a false assumption that goals set by evolution are somehow benign, and those that are self-set are not. The delegate gave the example of a pandemic flue that could wipe out millions of people around the world.

The panel thought the difference was one of timescales. While it is true evolutionary goals can awry, the speed that things can go awry with a superintelligence is so much faster.

SESSION 1: Creativity and Imagination in Humans and Machines

Chair: Professor Margaret Boden Research Professor of Cognitive Science, University of Sussex

Dr Demis Hassabis Co-Founder and CEO, DeepMind (Google)

Professor Simon Colton Professor of Digital Games Technologies, Falmouth University and Professor of Computational Creativity, Goldsmiths, University of London

Dr Demis Hassabis

"Intuition and creativity in the context of the AlphaGo match"

There is a long history of games research and AI. Ever since chess was "cracked" when Deep Blue beat Gary Kasparov, there has been the appetite in the AI world to take on the more complex game of "Go".

These are some of the characteristics of the game:

- It is more than 3,000 years old
- There are 40m active players; and more than 2,000 professionals
- It is played on a 19x19 grid
- There are 10^170 possible board configurations – more than the amount of atoms in the observable universe

That final point is crucial. It implies that, even if you were to take all the computer power on the planet and you ran it for a million years, it would not be enough to search through the different possible outcomes exhaustively.

> ***'The complexity of 'Go' makes the brute force search approach that Deep Blue took completely intractable'***
>
> ***- Dr Demis Hassabis***

With AlphaGo there were two main challenges: firstly, the search space is enormous[1]; also before AlphaGo it was thought impossible to write a sufficient evaluation function.[2]

Tackling these challenges produced two neural networks: the "policy network", which was developed by studying moves from 100,000s of recorded games; and the "value network", which predicts which side is winning on a scale from 0 to 1 on a turn by turn basis.

[1] For the average turn in "Go" there are 200 possible moves; in chess there are 20 possible moves.

[2] It is much harder for an AI to evaluate who is winning in a game of "Go" at any one time than it is in chess. Because chess is a destructive game (i.e. the game starts with a full set of pieces and the numbers reduce over time) the game simplifies as it progresses. "Go" is the opposite. Any evaluation of a mid-game position requires a degree of prediction.

In essence, by searching through the most likely moves, the policy network reduces the breadth of possible moves; and the value network reduces the depth of those searches. Before AlphaGo, the only way to evaluate a position was to use the "Monte Carlo Tree Search", which requires searching through millions of random play outcomes to find the best move. This is computationally very expensive and as such an inefficient method, but it was what took AI machines up from a very weak amateur level, to a strong amateur level. Its ability to make considered value judgements took AlphaGo a stage further.

First AlphaGo took on and beat the best other AI "Go" players,[3] then a professional human player and then the world champion, Lee Sedol. In a series of matches watched by more than 280 million people (more than the Super Bowl), AlphaGo won 4-1.

One of the greatest moves AlphaGo made during the contest was move 37 in game two; when the machine broke 3,000 years of history and tradition by playing a stone on line five, and then went on to take control of the centre of the board.

Professor Simon Colton

Computational creativity is the science, philosophy and engineering of computational systems which, by taking on certain responsibilities, exhibit behaviours that an unbiased observer would deem to be creative.

One project is called "The Painting Fool", which paints pictures of a subject, but the resulting image is, to an extent, a side-product. What is important is the behaviour of the software during the painting process, displaying imagination, intentionality and learning.

Another project, the "What If Machine", invented the original concept for a West End musical.

Within 20-30 years time we can expect iTunes to compose music; Google's search engine to write an entire magazine in response to a search term; and PowerPoint to add jokes to a presentation.

These advancements will be brought about by incremental improvements to software, not by a blurring of the lines between computational and human life forms. AI software does not pose an existential threat to humanity or human creativity.

Creativity is a secondary and an essentially contested concept. It is not an intrinsic function of a person or a piece of software, but something other people project into it. In some circumstances, if someone describes a person or piece of software "creative", it becomes so. But it is an entirely subjective judgement: in essence, if we are not arguing about creativity, we are not talking about creativity, and such arguing is a driving force for progress in society.

'What does it mean for software to be imaginative in a context where we don't pretend it's a person?'

- Professor Simon Colton

[3] AlphaGo won 499 out of 500 matches against leading AI players

The time has come to stop talking about software in human terms and to dispense with Turing-style tests, which can lead to encouraging naivety and pastiche production in software behaviour.

It is a bit of an insult to assume that the human race will slowly but surely engineer its own destruction. Instead, the development of creative AI technology will improve people's lives and provoke global tech firms to produce better, more intuitive programs.

One way to plan the development of creative software is in the following phases:

1. Software can generate content
2. Software can invent its own measures of value
3. Software takes ownership of its own creative processes and products
4. Software is able to write its own code
5. Software's creativity is taken so seriously that it itself can enter the debate over what creativity means

[The full text of Prof. Colton's talk is available here: http://metamakers.falmouth.ac.uk/rustat-talk/]

Discussion

Delegates discussed the implications of AlphaGo's search and evaluation functions. Could it assist in the formulation of scientific theories such as the quantum theory of gravity?

The panel said that it could. The ultimate goal in developing AI is to help scientists make bigger, faster breakthroughs. However, developing new scientific rules is still a long way off.

The main thing holding AI back from making great imaginative leaps forward is its complete obsession with 'the truth', which prevents it from spotting what element or elements are 'missing' from a discussion.

Another delegate asked how far the AlphaGo system could be used to play a game like Bridge, for example, where information is incomplete. The response was that the team is working on programs to play such games (e.g. multi-player poker).

Another delegate asked about AI's lack of a "body" and the impact of its absence on thinking.

The panel said, while AI developers take inspiration from psychologists, they are not programming brains. The emphasis at Deep Mind is on reinforced learning, which is a strong component in animal development.

But embodiment is vital if you are to develop a true cognitive system.

At Deep Mind, the AI models all use virtual agents or avatars, but they are treated as though they were physical agents whose knowledge comes strictly from perception.

The biggest roadblock on the path to achieving a superintelligence is going from perceptual understanding to abstract knowledge.

SESSION 2: *Human-AI Equivalence and Beyond: Personhood, Consciousness, Values*

Chair: Professor Murray Shanahan Professor of Cognitive Robotics, Imperial College London

Professor Tim Crane Knightbridge Professor of Philosophy, University of Cambridge

Professor Sarah Coakley Norris-Hulse Professor of Divinity, University of Cambridge

Professor Murray Shanahan

"The space of possible minds"

Biological minds comprise only part of the space of all possible minds. But what characteristics would non-biological minds have? We may have human-level artificial intelligences within 30-40 years, but that does not mean they will be human-like. Differences could include:

- Awareness of the world
- Cognitive integration[4]
- Awareness of the self
- Capacity for suffering
- Selfhood/autonomy

Intelligences can be ordered along two rough planes: their human-likeness and their capacity for consciousness. In the biological world, intelligences have strong positive correlation, with non-human-like things having a low capacity for consciousness, and more human-like things – i.e. mammals – having a greater capacity for consciousness.

But this is not true of AIs. It is easy to conceive of a zombie-like AI entity that is human-like, but with no interior life whatsoever; or a non human-like machine that has a level of consciousness that far outstrips our own.

Professor Tim Crane

When talking about human and AI equivalence, there are three questions that need to be asked:

1. What is it that computers actually do?
2. When considering possible developments in AI, what should we be concerned about?
3. Can computers be like us?

Any meaningful response to the third question requires one to ask: what are we actually like? And this is a question which is complex, multi-dimensional and still deeply disputed. For example, the questions of what intelligence and consciousness actually are have no settled answers. So how to even formulate the central question for AI is itself deeply controversial, for philosophical reasons.

[4] Defined as the ability to bring to bear the whole of the brain's resources to an ongoing situation

‘Machines will be capable, within twenty years, of doing any work a man can do’

- Herbert A Simon (1965)

The level of computer ability has improved drastically since AI pioneer Herbert Simon made his predictions in the 1960s. There is now more AI in a smartphone than Simon could have imagined. But no-one seriously thinks that these amazing machines are getting any closer to having general intelligence, let alone being conscious. When we seriously consider the question of whether we might torture a smartphone, for example, then we will be seriously asking about its consciousness.

Merely because some computers are very good at complex rule-governed games, or that they have abilities analogous to our own, does not imply that they are intelligent in the way we are, or that they are “like us”. The creators of Watson, IBM’s AI machine, claim that it learns “in the same way that humans do”: it “observes, interprets, evaluates and then decides”. But they also say that Watson can read “millions of unstructured documents in seconds”. Humans do not do that; so in what way is its learning process “the same” as ours? We need first to figure out how our intelligence and consciousness works, before speculating about what AI machines of the future can do.

Professor Sarah Coakley

Predictions about the future of AI are sometimes alarmist – even apocalyptic – in their moral implications. To quote US poet Louis Untermeyer: “Are we about to become the slave of what our slaves create?”

But while such fears need not be debunked or completely ignored, we should subject them to rational and moral investigation. In so doing we may find three reflections on the future of AI:

1. What meta-ethical picture is being presumed by AI fear mongers?

It is often assumed that finding the greatest good for the greatest number of people is what ethics is all about. Since a machine may be able to figure that out much quicker and better than we can, AI may soon be able to outmanoeuvre us ethically, making choices that are unpalatable to us.

But something has gone awry here. Such a consequentialist understanding of ethics abstracts us from a great number of features from our ethical lives: senses of fidelity and loyalty, for example, and our sense of community.

2. What if ethics is about the formation of virtue: how can we value-load machines?

If “learning to be good” is the project of a conscious, socially-related self, then how could AI replicate this in any meaningful way? In humans it is realised through a community of learning and a sensitivity to the thoughts and feelings of others. No such framework for machines exists.

3. Can machines do evil?

We can envisage a machine being destructive, but to call its actions “evil” is to attribute to it capacities for a richer moral meaning than it legitimately deserves. It would be more true to say that the (human) machine maker was more responsible than the machine for any evil doing.

As philosopher Nick Bostrom says: if we could solve the “value loading” problem of AI, we would still be confronted with the problem of which values to load. Until then we should focus on the ethical lives of the machine makers.

Discussion

A delegate said that the question: “can computers be like us?” is actually misleading. A more useful question might be: “can computers be like us in one specific way?” (Can computers be empathetic like us? for example).

Another delegate said that the term AI was too broad and does not reflect the granularity of its individuation. Sometimes we are talking about an AI that is a human-size entity; sometimes we are talking about a global, disembodied superintelligence.

This will have an impact on their senses of personhood, consciousness and ethical life. If the AI is some sort of overarching hive-mind, it will have a different set of sensory, ethical and conscious parameters than a human-like robot.

Another delegate said that anthropomorphism would have an impact on how human-like we perceive an AI to be.

A delegate said that because superintelligent machines could be developed that were not conscious, it could provide great insight into how unique certain “human-only” traits actually are. Perhaps creativity will be found to be not that mysterious after all, or intuition. Developing AI machines will be the best way to empirically test such constructs.

Some delegates discussed an AI’s ability to model the human brain. Perhaps quantum computers would be able to achieve this?

But another said that, while it is true we know how to model quantum systems, we do not yet know whether the brain uses coherent quantum states – and certain delegates felt that it might not.

Even if we could model the human brain precisely enough would that be the end of human identity and consciousness? The majority of delegates felt that it would not.

SESSION 3: *Ethics, Regulation and Governance*

Chair: Professor Huw Price Academic Director, Leverhulme Centre for the Future of Intelligence; Bertrand Russell Professor of Philosophy, University of Cambridge, and Co-Founder, Centre for the Study of Existential Risk

Dr Jatinder Singh Computer Laboratory, University of Cambridge

Professor Murray Shanahan Professor of Cognitive Robotics, Imperial College London

Professor Huw Price

Framing a discussion about AI solely in terms of risk – even existential risk – misses an important part of the story. In the case of a driverless car, the problem is to get us safely from A to B. In the case of the long-term future of AI, we don't know where B is – in other words, we have very little idea where the technology might take us. There may be many possible destinations, and as well as thinking about safety, we need to think about where we want to go. It may be a choice we only get to make once.

Dr Jatinder Singh

"Who do you sue?"

Liability – working out who is responsible – is complex in the algorithmic world. At the same time, it helps to regulate and shape the AI systems of the future.

Discussing liability with regards to machine learning is important because things will go wrong; harm will be caused; regulations and contracts will be breached. But an understanding of liability will also encourage a more responsible culture around AI creation.

A machine learning system entails an algorithm, which looks for patterns in data. It both learns from data and runs off data to produce outputs.

The key concerns are:

- The visibility and transparency of what's happening
- The degree to which ML systems can be constrained or controlled

If a machine learning system were hiring employees for example, we would need to ensure it was not being discriminative. So being able to see how the algorithm worked might reveal biases, these could be removed when noticed or a constraint be put on the algorithm to weed them out.

Data management is also relevant for managing responsibility – where the data comes from; its validity; and how it exudes. One clear example of this is the Microsoft Twitter bot, Tay Tweets, which began producing offensive outputs after it was fed on tweets from trolls.

'Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours'

- ***Telegraph headline 24/03/2016***

Though it seems one could not generally argue that "it was not me; it was just the ML system", there are complexities; e.g. concerning the ML supply-chain, the circumstances and workflows surrounding its use, and possible feedback loops between (directly or indirectly) competing ML systems.

Professor Murray Shanahan

"Technological unemployment"

What happens when technological improvements remove the need for jobs? One immediate example is self-driving vehicles. Self-driving trucks could threaten thousands of jobs in the US in the next few decades.

As technology becomes more sophisticated, so white-collar jobs will come under threat too.[5] In the future humans will have less work to do, so how should we manage this inevitable process? Maybe, like during the Industrial Revolution, new jobs will be created. But maybe they will not.[6]

Even if this helps create some form of utopian society – which relies on a universal basic income – we are left with another question: how do we want to live our lives?

Discussion

Delegates discussed the automation of jobs and the likelihood that "lucky" ex-workers would carry on receiving their pay for no work. Would commercial bosses allow it? If they would not, it was asked, what levers do governments have to regulate and coerce multi-national companies to treat ex-workers generously?

The panel said that developing cross-border regulatory frameworks are difficult enough as it is today, in reference to unemployment in industries such as the steel industry. With technological advancement, this will only get harder.

Another delegate quoted the feminist slogan: "the personal is political". If you wanted to update that slogan now, you would say: "the technical is political". But this is something rarely discussed, it seems. Part of the problem is that ideology is sometimes described as "what determines how you think when you don't know you're thinking." Because ideas about economy have become so ideologically engrained they are rarely examined or challenged.

Another delegate said that society needs massive redistribution of wealth so as to ensure that what is earned by machines does not just go to elites. That wealth should not go towards funding some basic minimum wage, but instead be used to set up dignified and worthwhile working

[5] E.g. Legal clerks, who search through case law to report to the barrister. The AI developed for Watson is very good at searching through unstructured documents and could do this job very efficiently

[6] *The Second Machine Age* by Brynjolfsson and McAfee is a recommended reading on this topic.

programmes for people – an enlarged public sector with jobs for carers, custodians and gardeners for public parks, etc.

A delegate asked if there was any evidence that new jobs would not be created when technology makes others obsolete. After all, they had done in the past.

A delegate responded that the pace of change is what sets this new wave of technological unemployment apart from all others. It will be very hard for people to respond to such rapid adjustment to the labour market.

But the same delegate proposed a solution. The key concept to concentrate on, he said, was to maintain "clean interfaces" within markets, ensuring competition works.[7] Competition will promote the right granularity of companies within an economy and curtail monopolies.

Another delegate said that curtailing monopolies and enlarging the public sector could both be achieved if people worked out how to tax AI properly.

One delegate said we should not be thinking about computers and humans as distinct; or artificial intelligence and human intelligence as distinct. We should be thinking about these things as hybrid, where human intelligence creates artificial intelligence and the two become different to and greater than the sum of their parts. Similarly, companies like Google and Facebook have desires and goals – to maximise shareholder value, perhaps.

So the ethical questions raised today are actually about already existing, fairly autonomous, fairly intelligent things. We may only not recognise them as such because we are looking in the wrong place – we expect them to be in boxes and they are not.

[7] An example of a "clean interface" could be the availability of data. Within a market like the EU there could be a rule that says: if you collect data in Europe you need to make it available to European companies at no cost

SESSION 4: *Anthropology and Theology in An Era of AGI*

Chair: Professor Andrew Briggs Professor of Nanomaterials, University of Oxford

Dr Timothy Jenkins, Reader in Anthropology and Religion, University of Cambridge

Dr Andrew Davison, Starbridge Lecturer in Theology and Natural Sciences, University of Cambridge

Professor Andrew Briggs

Can machines ever learn wisdom? If so, could they do it by reading ancient scriptures? Or Shakespeare? Could a machine believe or disbelieve in God?

Emotions are integral to the process of making wise decisions. But how can that be applied to AI? We know what it is to judge a person kind or an action kind, but what would that mean – if anything – of a machine?

Dr Timothy Jenkins

The response of the humanities to growing machine intelligence is a crucial one. As the humanities provide such a response, there are certain observations to bear in mind:

1. The humanities must show precaution

It is easy for the humanities to miss the significance of technological innovation. Philosophers and social commentators may criticise certain over-enthusiastic or overreaching scientific achievements but miss the "direction of travel". The humanities should not be too eager to classify what is going on. Instead it should be slow into the game.[8]

2. Traditional schemes of explanation cease to operate

Opposing notions (realism vs idealism, for example) cannot grasp the novelty of innovation. Incompatible schemes seem to be at work simultaneously at the forefront of AI development – ideas and materials are mutually implicated. But there is a strength here, namely:

3. The humanities is well versed in theorising these kinds of patterns.

Learning to describe 'the new' is central to the history and philosophy of science.

4. What kinds of problems might emerge?

These are not obvious. The normal ways of establishing order might fail. Successions in time fail; notions of cause and effect fail or become reversed; memory and prediction mutate. Even the identity of an actor with observations and motivations becomes multiplied.

[8] This perhaps accounts for why science fiction is nearly always wrong, in hindsight.

After category collapse, language takes on a different role. No longer a tool for designation, manifestation, and signification, language takes on active properties. Naming may call imagined entities into being, for example.

5. What areas of research should we look into in particular?

    a. The military – military innovations feed into domestic ones.
    b. The media – where the contemporary will be represented
    c. The social life of groups – which draws on these resources

6. The human mind will remain the measure, means and media of what is going on

The lesson is not that anything to do with artificial intelligence needs to be ruled out, but is that, regardless of the scale of AI advancement, it will be incorporated or embodied within human concerns.

Dr Andrew Davison

"What would a scholastic make of our discussion today; and what would he think was missing?"

Artificial intelligence – and perhaps consciousness – is the property of a newly emergent whole. Some may assume that theology regards the emergence of genuine wholes with their own integrity as impossible, even a threat. From a Christian perspective, however, this possibility presents no such threat. Christian theology is not inherently dualist when it comes to personhood or intelligence.

The openness among theologians to the emergence of such wholes is characterised by the many 13th-century theologians who staunchly advocated Aristotelian Hylomorphism. Hylomorphism proposes that a material thing amounts to more than what it is made out of—in other words, that the whole is more than simply a sum of the parts. On this view, the structure of the thing—the very coordination of the material that makes the thing what it is—is worthy of note, in addition to that material out of which the thing is made/structured.

In the context of artificial intelligence, this raises the question of whether a machine can ever be a coordinated "whole" in the same sense that a complex being like a human being is a "whole." For example, a computer that simply follows a program – no matter how elaborate – cannot be said to be imaginative, so it is hard to see that a new "whole" is created. But when we imagine more complex embodiments of AGI, we need to ask whether in such contexts a genuine co-ordinated whole has in fact emerged.

'If a sufficiently complex organic whole can be a thinking, conscious whole; then a sufficiently complex machine can be a thinking, conscious whole'

- **Dr Andrew Davison**

Every agent bestows something of its own form on what it creates. This idea is associated with the medieval theologian St Bonaventure, whose treatise “On the Reduction of Arts to Theology” is surprisingly relevant in certain ways to contemporary discussions of AGI. There he writes

about an artisan making an entity that could know and love its maker. This would be possible, St Bonaventure says, because of the similitude of the maker and the thing made.

If Bonaventure is right, then AGI will almost inevitably carry the “human stamp”, however advanced it gets. As a product of human imagination, manifestations of AGI will exhibit human perspectives, assumptions and values.

In that way, we should be more concerned about the ways AI will be like us, than we should be about how they will be different to us.

Discussion

Delegates discussed the relationship between the maker and the made. One aspect was whether it could be empirically deduced whether a machine was truly creating something or simply replicating what it had seen. Others wondered if the same question could be applied to Renaissance painters.

But a delegate said that when a machine creates its own code, it is truly producing something new because the original (human) maker does not know what is going to be created by its machine.

Another delegate asked how the position of God as ‘the Creator’ might be challenged. Humans are now creating machines that possess attributes of consciousness, when previously it was believed that only God could do that.

One response was that many theologians believe that the act of creation is a collaboration with God.

Others spoke about the fact that, while a superintelligence might surpass the capacity of a human mind in one function, humans have many facets to their intelligence, and it would not impact on the position of humans in the world.

There is something of the ‘humanness’ of humans that is impossible to replicate or reduce.

Contact

Rustat Conferences
Jesus College
Cambridge CB5 8BL

www.jesus.cam.ac.uk

www.rustat.org

info@rustat.org