Conference Report

The Singularity Summit

Imagination, Memory, Consciousness, Agency, Values

AI & the Future of Humanity Series
Science & Human Dimension Project
Jesus College Cambridge

26-27 September 2018

Contents

Conference Rapporteur: Rob Hart
Editor: Jonathan Cornwell

Jesus College Cambridge

SHDP

Science & Human Dimension Project
Jesus College, Cambridge

The Singularity Summit

Imagination, Memory, Consciousness, Agency, Values

AI & the Future of Humanity Series
Science & Human Dimension Project
Jesus College Cambridge

Day 1 Wednesday, 26 September 2018

11.00-11.30 Registration - Bawden Room, West Court, Jesus College, Cambridge
Tea and coffee served. Move to Frankopan Hall, West Court, for opening session by 11.30

11.35-11.40 Welcome - Frankopan Hall
Conference Chair: John Cornwell Director, Science & Human Dimension Project

11.40-12.10 Opening Talk
Definitions and models of the technological singularity
Dr Anders Sandberg Research Fellow, Future of Humanity Institute, University of Oxford

12.10- 12.40 Opening Talk
Explanatory and performance gaps in Artificial Super-Intelligence
Prof Steve Torrance Visiting Senior Research Fellow, COGS, University of Sussex; Professor Emeritus of Cognitive Science, Middlesex University

12.40-12.55 Discussion

12.55-13.45 Lunch - West Court Dining Room

13.50-15.10 Session 1 Epistemology
Machine Intelligence and the capacity for insight
Dr Marta Halina Lecturer in the Philosophy and Psychology of Cognitive Science; Project Director, Leverhulme Centre for the Future of Intelligence, University of Cambridge

The promise of analogy: talking about Machine Learning across interdisciplinary boundaries
Dr Andrew Davison Starbridge Lecturer in Theology and Natural Sciences, University of Cambridge

15.15-16.30 Session 2 Consciousness
Whatever mathematical thinking or understanding is, can it be the same as what a physical computing mechanism does when it is said to compute or 'understand' a mathematical function?
Antonio Ramos Diaz Centre for Metaphysics, Institute of Philosophy, University of Leuven

In defence of the possibility of artificial consciousness: computation, diagonalisation, and the qualia delusion
Dr Ron Chrisley Director, Centre for Cognitive Science, University of Sussex

16.30-17.00 Tea - Bawden Room

17.00-18.10 Session 3 Agency
Agency without Agents: If the role of consciousness can be reduced to the role of mental states, then the prospects for genuine AI seem very good. But if consciousness is irreducible, genuine AI may be impossible
Dr Markus Schlosser School of Philosophy, University College Dublin

Agency, memory and the cognitive ecology of AI
Dr Robert Clowes Universidade Nova de Lisboa

18.10-18.50 Free time

18.50-19.30 Drinks - Cloister Court

19.30 Dinner - Upper Hall, entrance off Cloister Court
Jesus College Bar in West Court will be open after dinner

Please note we will be filming the conference
@ScienceHumanDP

Jesus College Cambridge

SHDP

Science & Human Dimension Project
Jesus College, Cambridge

The Singularity Summit

Imagination, Memory, Consciousness, Agency, Values

Day 2 Thursday, 27 September 2018

08.30-09.10 Tea and coffee served - Bawden Room, West Court, Jesus College, Cambridge
Move to Frankopan Hall for first session by 09.10

Chair: Jonathan Cornwell Executive Director, Science & Human Dimension Project

09.15-10.30 Session 4 Identity - Enactivism - Embodiment
Embodied perception and action - a challenge for AI?
Dr Simone Schnall Reader in Experimental Social Psychology and Director of the Cambridge Body, Mind and Behaviour Laboratory, University of Cambridge

RECtifying Intelligence: Radical Enactive/Embodied approach to Cognition and Intelligence
Prof Erik Myin Professor of Philosophy, Director, Centre for Philosophical Psychology, University of Antwerp

10.30-10.55 Break - Bawden Room

10.55-12.15 Session 5 Identity and Self - Panel
Someone but not a Self
Dr Daniel De Haan Research Fellow, Ian Ramsey Centre for Science and Religion, University of Oxford

Corporate individualism, robots and AI: a call to Martin Buber's I and Thou
Prof Kathleen Richardson Professor of Ethics and Culture of Robots and AI, De Montfort University

Robots with families: on identity and organic continuity
Fr Ezra Sullivan OP Faculty of Theology, Angelicum, Ponitifical University of St Thomas Aquinas, Rome

12.15-13.05 Lunch - West Court Dining Room

13.05-14.20 Session 6 Ethics
Brain-Computer Interfaces: new varieties of cognitive imbalance
Dr Karina Vold Faculty of Philosophy; Leverhulme Centre for Future of Intelligence, University of Cambridge

Ethics for E-persons
Prof Steve Torrance Visiting Senior Research Fellow, COGS, University of Sussex; Professor Emeritus of Cognitive Science, Middlesex University

14.25-15.40 Session 7 Aesthetics and Imagination
Should an AI cause insult and harm?
Prof Simon Colton Chair in Computational Creativity, Queen Mary University of London

Imagination and memory
John Cornwell Director, Science & Human Dimension Project, Jesus College, Cambridge

15.40-15.55 Closing comments and feedback

15.55-16.30 Conference close and tea - Bawden Room

Please note we will be filming the conference
@ScienceHumanDP www.science-human.org j.cornwell@jesus.cam.ac.uk

Jesus College Cambridge

SHDP

Science & Human Dimension Project
Jesus College, Cambridge

Science & Human Dimension Project - The Singularity Summit - Executive Summary

This conference probed more deeply into the technological singularity, examining the philosophical, psychological and ethical issues that arise out of the numerous research projects going on around the globe that seek to emulate and exceed human intelligence, or at least the products and abilities that stem from it.

The conference talks were grouped broadly into seven themes:

1. Epistemology
2. Consciousness
3. Agency
4. Identity - Enactivism - Embodiment
5. Identity and Self
6. Ethics
7. Aesthetics and Imagination

Summary of talks

Anders Sandberg sought to provide clarity on the term singularity, revealing it to be a "composite" term rather than a singular entity. He notes that common usages of the term refer to rapid acceleration, points at which our understanding breaks down and notions of self-improvement.

Steve Torrance outlined the risks of performance gaps in both current and potential future AI systems. Torrance suggests that a number of gaps might be "in principle gaps", stemming from fundamental failures in our theories of mind, in particular the computation theory of mind that seems to dominate scientific research today. The deficiencies of the computational theory of mind were a recurring theme in both talks and discussions throughout the conference.

Andrew Davison looked at the language we employ when talking about AI, making note of the talking past one another that can often occur when people in different disciplines get together to talk about AI. In particular, he suggests an alternative, middle path to speaking univocally (using the same word in different settings with the same meaning) and equivocally (using the same word in different settings to mean different things): analogy.

Marta Halina took a more targeted approach, examining whether AlphaGo could have exhibited creative insight. Drawing on studies of human and animal cognition, she concludes that it is unlikely AlphaGo would meet such criteria.

Antonio Ramos Diaz examined the nature of mathematical understanding, drawing upon the unpublished works of Kripke to provide an additional case against the computational theory of mind which supplements the commonly given Gödelian argument from Lucas and Penrose. The argument asserts that whatever fixes the formal properties of a function cannot be accounted for by intrinsic properties of that function. In humans such an extrinsic factor is represented by the human mind.

Ron Chrisley rejected the argument from diagonalisation, claiming that it is not required to compute the exact same functions in order to simulate something, say consciousness, perfectly.

Markus Schlosser examined agency and consciousness. He looks to Eastern philosophies to ask if agency might still exist without an agent or the self. From this perspective, he concludes, the very idea of a conscious mental state is a misconception.

Robert Clowes looked at agency from a different perspective, arguing that the current connected ecosystem of technology could actually enhance our agency, rather than diminish it, as many argue. He does, however, note the danger posed when we do not know or have access to the technologies, such as social media algorithms, that govern our lives.

Simone Schnall looked at the issue of embodiment in detail, noting studies that show how closely intertwined perception, affordances for action and the body are. This is a problem the AI community will need to face one day, noting that current efforts lack embodiment and affordances.

Erik Myin presented an alternative view, arguing that the vast majority of cognition is contentless. Ultimately, cognition, including perception, is intentional without involving contentful representation or computation. Any computational cognition must be scaffolded with another faculty like language.

Daniel De Haan took a closer look at what we mean when talking about ourselves, claiming that there is an important distinction between us as the self and us as a someone. Importantly, he says, in order to answer the question of 'who am I' we must adopt the stance of a someone, an individual in constant development.

Kathleen Richardson examined the work of Martin Buber, to whom the word I is a composite of I-it and I-thou, contrasting this with Cartesian ego centrism. Interpersonal relationships are crucial to our understandings of ourselves, and we will need to think about how to incorporate this view into our developments of AI.

Ezra Sullivan observed the fundamental differences in family that exist between AI and biological life. Lacking a place in this 'family of living', Sullivan argues that there might be an unbridgeable gap between any AI identity and that of humans or other living things.

Karina Vold looked at the fraught ethical terrain of brain computer interfaces (BCIs). She laid out the terrain of possible devices, paying particular attention to the issues enhancement raise. Vold noted the risk of hacking or jacking of the brain, but is particularly concerned about the potential for enhancement to throw off natural cognitive balances that exist in the brain.

Steve Torrance looked at the potential issues that arise from electronic personhood, the likes of which are being discussed in the EU. Particularly, he wonders whether a notion of legal personhood might engender some secondary moral status on such electronic persons.

Simon Colton questioned whether AI can genuinely cause insult and harm. In particular, Colton notes that AI, lacking authentic life experiences, may never be able to truly cause insult. He points to

ways in which we might develop the lived authenticity required to achieve such a task.

John Cornwell asked "how do you feel about AI?" Cornwell admits that while he does share the optimism of many, he also "feels a sense of increasing melancholy and angst along with the optimism."

Key points raised

Limitations: A recurring theme in the conference was the limitations of the computational theory of mind. Such limitations could lead to performance gaps, says Steve Torrance in his first talk.

Embodiment: Many participants spoke in favour of more embodied approaches to cognition rather than what appears to be a reliance on computational theories of mind. Murray Shanahan notes the important role embodiment plays in his work, and Simone Schnall points to many scientific studies that support embodiment. "One has to look at the body to fully understand cognition," she says. Marta Halina agrees, saying she is "very much in favour of embodied cognition". Anders Sandberg, in a comment detailing how cars and other objects we use extend our senses beyond our bodies, wonders what might happen when such extensions of ourselves have a degree of autonomy.

Linguistic difficulties: A number of participants noted the inherent linguistic difficulties in interdisciplinary discussions about AI. "Conversations can be quite fraught," said Andrew Davison. There is also a great deal of talking past one another. Davison believes centuries of scholastic thought on the nature of analogy would be useful here. William Clocksin, however, points out that today we have certain distinctions that the scholastics would not have had. This might limit the use of that thought. One participant questioned the value in changing the language we use to talk about AI, asking whether it would actually be useful to go into a room of people developing driverless cars and saying they need to overhaul their language.

Today's AI vs AGI: A number of participants stressed the need to make a clear distinction between the AI being developed today and the possibility of General or Superintelligent AI. "We should be making a clearer distinction between the possibility of artificial general intelligence and artificial intelligence that's actually being built today," said one participant. Simon Colton went as far as to apologise. "As a scientist I want to apologise for the fact that… we've let the impression the rest of the world has about AI get out of hand… we as scientists should be out there more explaining what actually is the state of the art and that it's not really at the stage where we should be concerned." Murray Shanahan voiced a similar opinion, observing the fact that many scientists are not trying to emulate the human mind but are instead trying to create systems with sophisticated abilities. Simone Schnall said that it's "almost impossible to not take the human being as the default." Participants expressed divergent opinions on the use of science-fiction in discussions, some believing it a useful tool, others a distraction.
Anders Sandberg offered a countering opinion. He notes that while many predictions and models of the singularity have failed in the past, "uncertainty does not mean we should be complacent."

Social and Political Aspects: In a similar vein, many participants stressed the need to, in the words of Kathleen Richardson, "bring it back to social reality… If we're talking about singularity…

what is going on politically in our world that has got us to this point where we're thinking about this idea of merging with machines?"

Ezra Sullivan dedicated his talk to the family of interpersonal relationships amongst all living things, and an AI's place within this nexus. Neil McBride also made such a distinction, saying he believes there are two areas we should be paying attention, the political and social possibility of the singularity and the scientific possibility.
Fenella Cannell spoke similarly of the importance of social context: "the fantasy of the asocial person is a particularly dominant and tempting error of the contemporary West." We must be careful not to project this model of humanity when engaging with the development of AI. The often hidden social aspects of power at play in AI systems were also raised as a point of concern. This was raised by Robert Clowes at the end of his talk, and echoed by a later commenter.

Inclusivity: The need for greater inclusivity and diversity in AI development was raised by many participants. "Particularly women," says Kathleen Richardson. Daniel De Haan agrees, stressing the need for academics to engage with others outside their disciplines or risk becoming 'siloed'.

Contentful Representation: Erik Myin claimed that cognition, including perception, may be intentional without contentful representation. Marta Halina and Ron Chrisley both expressed their dissatisfaction with this idea. Halina says she's never really "been on board with antirepresentationalism… I have trouble seeing how it can account for cognition and bodies in the world." Chrisley asserts that representationalists have "met the burden of proof", and states that antirepresentational views need to do the same to be taken seriously. "It risks being a straw man," says Sally Davies.

Agency: Markus Schlosser turns to Eastern philosophy to provide a new perspective, asking whether we might have agency without agents. It is a theme that can be found in many Eastern philosophies, though it left many participants dissatisfied. David Miller commented that "there are other avenues we could take" that do not "dissolve people." Another commenter said that a lot of explanatory power is lost when removing the agents.

Machine Consciousness: The possibility and potential nature of machine consciousness proved contentious. Antonio Ramos Diaz presented an argument against its possibility in his talk, and Ron Chrisley dismantled common arguments against its possibility in his. One participant mentioned a "certain anxiety about tipping over into consciousness (that) comes from a capacity to suffer" in his work.

Ethics of Enhancement, Brain Computer Interfaces and Implants: Karina Vold spoke on the ethics of brain computer interfaces, particularly the issue of cognitive imbalances we might encounter. Sally Davies noted the complexity of the situation given the incommensurability of values pre and post implantation. Sumit Paul-Choudhury noticed a shared ethical terrain with chemical cognitive influencers, and a number of participants, such as Michael Stevenson, observed the potential for enhancement to threaten principles of equity (in Stevenson's case this referred to education.)

Human vs AI Potential: Throughout the conference, participants emphasised the need to think

about human potential, not just AI's. "I'd love to see us concentrate more on potential within human intelligence" than on machine intelligence. Yaqub Chaudhary stressed the importance of children's interactions with AI, and Michael Stevenson also observed the challenges AI posed to education. He says that policymakers around the world are doing the right thing in addressing what "children should learn to nurture what's distinctly human." Bonnie Zahl echoed this sentiment, this time with regard to raising children. "If there's one thing I can do better than a machine it's that I can be a better parent."

Beth Singler and Alicia Perez neatly captured themes that unified talks and discussions throughout the conference. "AI is pushing us to "reidentify" what it is to be human, says Singler. "We've started to move into a void space where we are less certain than before." Perez thinks "artificial intelligence might be a natural evolution of being human."

Singularity Summit - Science and Human Dimension Project Report

This two-day conference—the singularity summit—is the fourth of five conferences designed to bring together often disparate disciplines to discuss advances in artificial intelligence (AI). The conferences are held under the auspices of the Science and Human Dimensions Project (SHDP), based at Jesus College, Cambridge, which was founded in 1990 as a means of enhancing the public understanding of science. Three of these conferences, this one included, are made possible through the generous support of Templeton World Charity Foundation (TWCF).

The first of these conferences, held in June 2016, explored what kinds of questions those in the arts and humanities should be asking of experts in AI. The second, held in September 2017, brought scientists from Google DeepMind together with scholars in the arts and humanities in order to explain and make better understood the science behind how their machines work. The third, held early in 2018, examined artificial intelligence through the lens of science fiction, and the predictions and visions therein made. The final and fifth conference, to be held in 2019, will focus on the ethics of AI systems.

This conference, as its name suggests, will probe more deeply into the technological singularity, examining the philosophical, psychological and ethical issues that arise out of the numerous research projects going on around the globe that seek to emulate and exceed human intelligence, or at least the products and abilities that stem from it. In most instances, such projects occur in isolation, distant from the discussions occurring in humanistic and artistic disciplines that would both benefit from and be of benefit to these ongoing scientific projects. With that in mind, the conference will, over the next two days, explore and address the questions AI and the singularity raise relating to themes like embodiment, identity, agency, imagination, free will enactiveness and creativity. What are the prospects for machine consciousness? Will AI alter our understanding of what it is to be human? How might AI go about solving problems, what agency might it have and how will it go about making value judgments usually reserved for human minds alone.

The conference talks, beyond those introductory ones following immediately, are grouped broadly into seven themes:

1. Epistemology
2. Consciousness
3. Agency
4. Identity - Enactivism - Embodiment
5. Identity and Self
6. Ethics
7. Aesthetics and Imagination

Conference Introductory Talks

1) Definitions and models of the technological singularity. Dr Anders Sandberg, Research Fellow, Future of Humanity Institute, University of Oxford.

The concept of the technological singularity is one encountered frequently in discussions around

AI. For a term so commonly used, it lacks precision and clarity, says Dr Anders Sandberg of the University of Oxford's Future of Humanity Institute. Generally when people bring up the Singularity it's assumed we all understand what that means. This allows people to elide a lot of meaning and do a lot of sneaky things in their papers and arguments.

Rather than representing some single idea or entity the Singularity is better thought of as a composite idea people pick and chose elements from. In name, the term can be traced back to science fiction author Vernor Vinge. As an author of science fiction, Vinge was acutely aware of his inability to write about superintelligent characters or the world they inhabited. It would be incomprehensible not just to the reader but to his own abilities to predict. So he drew on the term from mathematics, whereby a singularity is a point where a function is undefined in some suitable sense.

But Vinge was far from being the first or only person to be thinking about what we would now regard as the singularity; we can point to the thinking of figures like Von Neumann, Turing, Good and Kurzweil who all feature elements we would recognise as the singularity in their thinking.

The question of who came up with the concept is a trick question. They are not really the same concept but things that appear similar when placed together. That said, a few trends do emerge. One of the most common usages of the term is to refer to accelerating change, and another to that of self-improvement. Some refer to the singularity as more of a phase transition in society, or as some special inflection point in history. Some go as far as to argue that the singularity represents a genuine run off to infinity, as in mathematics. In some instances, the Singularity has become almost like a religion: the rapture of the nerds imagining that intelligence is going to take off and we're all going to some digital Nirvana.

And with many different definitions come many different ways to try to model the singularity. Generally these models tend to be very large in scale than fine grained. The models vary in size and scope, but generally speaking we should treat them with scepticism. Historically speaking, the evidence suggests that those who are experts today are unlikely to be in the future, and uncertainty clouds many of these predictions. Nonetheless, uncertainty does not mean we should be complacent.

2) Explanatory and performance gaps in artificial super intelligence. Prof Steve Torrance, Visiting Senior Research Fellow, COGS, University of Sussex; Professor Emeritus of Cognitive Science, Middlesex University

Philosophically speaking, discussions of AI tend to go hand in hand with the computational theory of mind. Such a view about the nature of the mind explains mental functions in terms of computation and in terms of brain centred information processing terms. However the computational theory of mind is not the only theory that could underpin AI, and there are other embodied approaches that are discussed later in this conference. If the supporting theories that underpin AI aren't adequate then it may leave gaps in performance - that's something we need to look at.

Such performance gaps could be trivial, like many we see today, for instance in failing spellcheckers. On the other hand, as systems grow more powerful and more complex, it is possible

that such performance precipices could become severe and even dangerous.

And there are indications today that the computational theory of mind may already be insufficient to underpin our understanding of AI. Without a digestive system, how can an AI fully be said to understand anything relating to food or our human experience of it? What about the kinds of interpersonal connections and relationships humans form? And naturally with such theoretical deficiencies come performance deficiencies. These can be referred to as in principle performance gaps. How can a robot provide, say, care if it does not and has never experienced such needs for itself, Torrance asks, referring to an earlier lecture by Margaret Boden examining the topic.

Ultimately, we should be seriously considering the possibility of these in principle performance gaps, and that our theories of AI may be deficient, before we race ahead with projects to create a general or super AI. However we define it (AI), we want it to have as few performance precipices as possible and, ideally, if they do occur we want them to be known to the system and designers. Given the potential for AI to radically reshape and alter society, it's important to begin such discussions now.

Concluding Comments, Discussion and Recommendations from Participants

- Dr James Dodd, a Rustat Conferences member, observes that many of the great advances in "physics over the last century have been done entirely without the assistance of computation". Computers were, effectively, relegated to the status of observational tools like microscopes. He goes on to suggest that human intelligence is now limited and defined by a search for a quantum theory of gravity. He wonders whether machine intelligence might be applied to this endeavour. Sandberg responds, saying that the reason computers do not play a huge role here is due to the desire for explanations. "We would be relatively unhappy with Einstein if he just gave us reams of papers with very good predictions about astrophysics but no good explanation. This is of course what we are complaining about now with neural networks."

- The Rev'd Dr Tim Jenkins suggests that many of the rapid technological developments we have seen in history, that are our closest approximations to singularity, have come about because of war. "If you want something that looks like a singularity you just need a major war," he says. Regarding controls and regulation, Jenkins suggests we should be looking to into legislation.

- Lord Martin Rees made three comments contrasting comprehension and computing. A simple example might come from coordinate geometry; we might comprehend the formulae for discs, but a human will never comprehend the complexity of the Mandelbrot set even though it is relatively easy to compute. Second, Rees looks to fluid mechanics, where we might compute to astounding degrees of complexity but we will still lack the kind of insight that comes at higher levels. Third, Rees returns to quantum gravity theories. He believes it's conceivable for string theory to be correct but we would never be able to do the mathematics well enough to prove this.

- Another commenter would "love to see us concentrate more on potential within human intelligence" than on machine intelligence. She goes on to add that there are so many blocks to achieving the potential of human intelligence, many of which could derive from social, say schooling, or bodily, like nutrition, factors. "I'd like to see us address that as fully as we're addressing machine

intelligence." In response, Sandberg notes the propensity to refer to AI as a singular discrete entity, and points to other areas such as intelligence amplification that might be used to achieve and elevate human potential.

- Neil McBride from De Montfort University believes there are two areas we should be paying attention to, one being the scientific possibility of the singularity and the other the social and political possibility.

- Murray Shanahan, of Google's DeepMind and Imperial College, London, wanted to emphasize the importance of embodiment in his book. More generally, he notes that most people working in the field aren't interested in building models of the human mind but instead systems with sophisticated abilities. "They may not resemble human minds or human psychology at all," he says.

- Another commenter believes attempts to emulate the mind to be misguided, just as studying bird physiology is for achieving flight in history. Instead we need to be hunting for these underlying principles, just as an understanding of aerodynamics allowed us to achieve flight, albeit in a completely different way to birds.

Epistemology

When discussing artificial intelligence, one comes across a great deal of shared vocabulary in common with studies in psychology, the arts and humanities and other disciplines. Terms like creativity, insight, and even intelligence itself all come packed with a great many assumptions, values and understandings. But do we mean the same thing as we do in other disciplines, such as psychology, when we describe a computer as creative or intelligent? In this section, Andrew Davison, the Starbridge lecturer in Theology and Natural Sciences at the University of Cambridge, examines what kinds of meaning exist when using these shared terms. In a more targeted approach, Marta Halina, a lecturer in the Philosophy of Cognitive Science at the University of Cambridge and programme director for the Kinds of Intelligence project at the Leverhulme Centre for the Future of Intelligence, looks to contemporary research in cognitive psychology, particularly that taking place on non-human animals, to see if this sheds light on the nature of terms like insight and whether computers can have them.

1) The promise of analogy: talking about machine learning across interdisciplinary boundaries. Dr Andrew Davison, Starbridge Lecturer in Theology and Natural Sciences, University of Cambridge

As this conference illustrates, there are remarkable opportunities to be had when bringing together scientists working at the forefront of AI and scholars in the arts and humanities. Discussions can be immensely productive, but they can also be incredibly difficult and one often has to work hard to find common ground between disciplines. In no small part, conversations are so hard because of the tensions that arise when having to use old and existing vocabulary in new, and perhaps unsuitable, circumstances. Conversations between domains can be quite fraught. Moreover, there is a risk that we are talking past one another. Conversations had at previous iterations of the SHDP conferences at Jesus College are cases in point, conversations which were "not always easy".

Centuries of Scholastic thinking on language might, then, offer us ways of thinking that could "open up some complex and ambitious avenues for exploration" and deeply enrich our discussions surrounding AI. A scholastic would identify two broad camps for claims made about AI today. On one hand, we have univocity, whereby we use a particular word or phrase in different settings but mean the same thing by it. The use of the word "college" in talking about Jesus College and Emmanuel College, for instance, is intended to mean the same thing. With respect to AI, Davison continues, this can often be seen in the "characteristically bold claims" some scientists make for their systems, systems that can "see scenes", "intuit Newton's laws," or "understand, remember or show creativity."

On the other hand, we have what scholastics call equivocity. Equivocity is where the same term is used in different circumstances intended to have very different meanings. For instance we have the "bark of a tree and the bark of a dog" or "the bank of a river and the bank in the high street." When it comes to AI, then, we might discuss memory or understanding or creativity but the term has a divergent meaning to how it is used when discussing, say, humans or animals. Equivocity is commonly used by those in the arts and humanities when discussing AI.

When it comes to AI, neither equivocity nor univocity are entirely adequate. Speaking equivocally may undersell some of the astounding capabilities some AI systems display, while speaking univocally may go too far in the other direction, overselling what there is and claiming too much. All in all , equivocity and univocity slugging it out does not make for good communication. We need an intermediate path.

Analogy offers a different way of speaking about AI. It recognizes similarity and acknowledges difference. The similarity is all the more striking given the difference. "I think that's what we're seeing with artificial intelligence".

There are centuries and centuries of scholarship on the nature and types of analogy, and aside from providing a useful linguistic tool with which to talk about AI, it also gives us the ability to discuss some of the more metaphysical issues surrounding AI, such as exploring the relationships between the two entities under discussion. One possible avenue of exploration is the relationship of making, which might lead us down interesting theological areas of study. Another is of causation, which might lead to discussions of a creator, or that the environment affords a certain possibility.

Concluding Comments, Discussion and Recommendations from Participants

- Kathleen Richardson of De Montfort University believes that it's correct to be thinking more critically about our language with AI, but wants to "bring it back to social reality." She notes that throughout history the science, as was the case with Malthus and Darwin, has been informed by the political ideas of the day. "If we're talking about singularity… what is going on politically in our world that has got us to this point where we're thinking about this idea of merging with machines."

- Iason Gabriel, from DeepMind Ethics and Society, wonders whether another way of interpreting the dissonance at the start of the talk would be to say there are two groups of people, one saying there is nothing unique about humans and the other, the technologists, worrying that there is something unique about humans that could not be replicated. One group of people are afraid about

uniqueness and another are afraid of letting it go. "There's always going to be an element of anything that we do that comes from what it is that our culture helps us to think as being thinkable," replies Davison.

- Fenella Cannell, an anthropologist at the LSE, added "I didn't entirely see the interactions in just the two contrastive ways you described," she says. Comparisons very much depend on who are making them and the powers they might levy. "You have to think about the flows of power which are selecting what comparisons can be made and how they can be communicated at particular points in our social history. They are also prohibiting certain comparisons from being made." There is no definition of humanity that would be apolitical.

- William Clocksin of the University of Hertfordshire believes we have certain distinctions available to us today that the scholastics would not have had at the time. "It goes much farther beyond the bark or the dog or the bark of a tree. Instead, the distinction may be the bark of the dog and the bark of the recording of a dog… The question is whether there was the intellectual apparatus in the scholastic times to be able to establish a distinction about reasoning about that kind of thing."

- Ron Chrisley wonders whether some of these notions do "precisely apply" univocally. He asks whether there are boundaries we might have to help us decide on these relationships.

2) Machine intelligence and the capacity for insight. Dr Marta Halina, Lecturer in the Philosophy of Cognitive Science; Project Director, Leverhulme Centre for the Future of Intelligence, University of Cambridge

For a long time, chess was viewed as the "drosophila of AI". But, a more appropriate model might be the ancient Chinese game of Go. Contrasting to chess, Go cannot simply be solved by superior computing power, its board possessing more permutations than are possible to compute. Indeed it is "so complex it is often described as requiring human intuition to play." As it is impossible to fully calculate and compute moves to rationally make decisions, players "instead have to rely on what feels right or what looks good… professionals often describe it that they're relying on their intuitions."

With this creativity in mind, I hope to explore the question of "whether AlphaGo is capable of creative problem solving," drawing on studies in human and animal cognition to do so.

AlphaGo, a product of Google's DeepMind, broke records and expectations by besting champion Go player Lee Sedol in 2016. It was a milestone many thought years away. In the commentary of the match, and repeatedly since, a number of moves AlphaGo made have been described as 'beautiful' and even 'creative' by Go experts and DeepMind officials. Indeed, the machine itself recognized it played moves that were unlikely to have ever been played by a human player in a similar position, and it is possible that such innovative play was decisive in the machine's victory.

But what exactly is meant by creative or beautiful in this context? Insight might be a better word, being more common in studies of animal and human cognition. In these fields, several variables—a

novel situation, no prior experience of the problem, the behaviour exhibited must be valuable in some way to the task at hand, and a demonstration of a deeper, causal understanding of the task—are used to determine whether something has been creative or insightful. Halina points to New Caledonian crows bending sticks in order to acquire an otherwise unreachable object in experiments. The animal appeared to demonstrate creative insight, though on further observation in the wild this was rejected; the crows had a habit of regularly bending twigs when they encounter them, meaning the behaviour was not novel.

So how does AlphaGo fare with these criteria in mind? The task—of playing Go—was not novel to the machine; AlphaGo played millions of training games either with itself or programmers during its development. Many more games, actually, than human players. One could also argue that given that much experience it is unlikely it has a deep understanding of the game, but rather has encountered many more situations than Lee Sedol has. However, it is important to note here that the opacity of the AlphaGo system makes it difficult to "know what it's representing." In short: we can't tell what its understanding is. That said given the significant retraining AlphaGo would require to play variants of Go, such as on a 3D board, retraining which human counterparts would not require, it could be argued that AlphaGo does not demonstrate a deeper understanding of the task.

Ultimately, though, there's "no underlying theory" how animals do or do not exhibit creative insight. "That's something I'd like to develop," says Halina. Philosopher Daniel Dennett "talks about different types of learners and that's one general way to characterize the differences we're getting at here." One might look to hypotheses for counterfactual competence and analogical reasoning. "The idea here is that you have a model of the world and specifically the generative structure of the world."

Concluding Comments, Discussion and Recommendations from Participants

- Erik Myin, from the University of Antwerp, asks whether the birds exhibit counterfactual competence, how they might have acquired it and whether this is required for Skinnerian learning. Halina, in response, points to sceptics who continue to give alternative explanations whenever evidence is put forward suggesting deeper understanding of psychological states in, say, chimpanzees.

Consciousness

Whether or not artificial intelligence, particularly general or superintelligent AI, might one day be considered conscious is a particularly common trope in works of science fiction, and one that occasionally makes it into current, less-fictional discussions of AI today. Even a remote prospect of artificial consciousness raises a series of unique ethical, social and philosophical quandaries, quandaries that should be addressed beforehand, lest we run the risk of exploiting and potentially harming another conscious entity. With these concerns in mind, an active and contentious area of study relates to whether consciousness could in principle be observed in machines, and if so what form it might take.

1) Whatever mathematical thinking or understanding is, can it be the same as what a physical computing mechanism does when it is said to compute or 'understand' a mathematical function? Antonio Ramos Diaz, PhD Candidate, University of Leuven.

Some of the most common tasks computers and machines are set pertain to computing mathematical functions and undertaking analysis. In many areas of mathematics, machines outpace their human counterparts by leveraging vastly superior powers of computation, and in general machines are now capable of mathematical undertakings beyond the reach of the human mind. But what is the nature of this mathematical 'thought'? What does it mean when a physical computing mechanism is said to compute or 'understand' a mathematical function, and is this meaning the same as when we talk of human mathematical understanding? Ultimately this "debate is whether consciousness can be duplicated rather than simulated."

The most famous argument put forth against the duplication of formal mathematical understanding by computers is the Lucas-Penrose argument. The argument, put simply, utilizes Gödel's famous Incompleteness Theorem to demonstrate the falsity of the computational theory of mind; the argument asserts that it is, as a matter of principle, impossible to duplicate the human mind and its many capacities in a machine.

There is also, Diaz says, another argument to be put forth against the computational theory of mind, this time derived from the unpublished work—lectures, in this instance—of the philosopher Saul Kripke. Kripke's argument is stronger, and "levels a more fundamental attack on computational accounts of mathematical and logical understanding and stands or falls independently of any Gödelian argument." The argument states that whatever fixes the formal properties of a particular function cannot be completely accounted for by the intrinsic properties of the function; extrinsic factors will be required which, in the case of humans, represent the conscious mind. For computers, such an extrinsic factor needed to bridge between mathematical and physical ontological categories, is represented by the intentions of the computer's designer. "Kripke's point (is) namely that the physical structure and causal makeup of a mechanism alone do not suffice to determine which function a system computes. If (a function) gamma fell from the sky, says Kripke, with no instruction manual whatsoever there would be no way of telling which function (of two, an identity and a counter function) the system computes." It therefore follows that the machine cannot be considered conscious, and, when speaking of mathematical and logical understanding, a computer is doing something different compared to a human undertaking the same process of logical and mathematical understanding.

Concluding Comments, Discussion and Recommendations from Participants

- John Cornwell, Director of the Science and Human Dimension Project, wonders whether Lucas' paper would have been quite the same if he had been exposed to AlphaGo and its intuition.

- Ron Chrisley points to two possible approaches that haven't been considered: 1) the so called natural function, is that when in a survival context, much like Dawinian evolution, an AI system might come up with a "natural function that an artificial system might come to have through its own adaptivity." 2) an interpretationist approach, argues that there are many alternative computational descriptions to understand your computer right now but there is one that would stand out as a clear

answer.

2) In defence of the possibility of artificial consciousness: computation, diagonalisation, and the qualia delusion. Dr Ron Chrisley, Director, Centre for Cognitive Science, University of Sussex.

Beyond the Lucas-Penrose argument, as well as that from Kripke outlined in Antonio Ramos Diaz's talk, Ron Chrisley from the University of Sussex points to other arguments that are often put forth against the possibility of artificial consciousness, exploring one in particular, itself a more generalised version of Penrose's above: diagonalisation. The diagonalisation argument claims that, given the existence of questions relating to computation that can only be answered by (conscious) humans, a machine can never fully behave like a conscious being. It then follows that, given the inability to behave in the same manner as a conscious human being, a machine or computer cannot be conscious.

Chrisley, in his talk, rejects these arguments, maintaining that neither give us adequate reason to doubt the existence or possibility of a conscious computer. Regarding diagonalisation, Chrisley explains, even if one accepts the premise of humans being able to answer questions about computation that computers cannot, the rest of the argument does not necessarily follow. In order to perfectly simulate something, say the aforementioned behaviour indicative of consciousness, one need not compute the exact same functions. "Even if there are questions that we can answer correctly that no Turing machine can, this does not mean that Turing machines cannot generate the same behaviours as we can. Even if that were true it doesn't follow that behaving like a conscious human is a requirement for being a conscious being. It might be that computers could be conscious in a non-human way even if they can't be conscious in the same way that humans are."

Concluding Comments, Discussion and Recommendations from Participants

- A speaker asked Chrisley to clarify his distinction between human and non-human consciousness which was mentioned in the talk. Chrisley notes that it could share some properties, such as awareness, but would be in non-humans, like animals. "People usually don't make the distinction… maybe they just mean simulated human machine consciousness is impossible."

- John Cornwell raises the question of platonic mathematics with relation to both talks, especially as it is a key theme in Penrose's work. Chrisley feels platonic work might be reformulated to work in this context, but notes he focused more on what computation is for his talk.

Agency

1) Agency without agents: if the role of consciousness can be reduced to the role of mental states, then the prospects of genuine AI seem very good. But if consciousness is irreducible, genuine AI may be impossible. Dr Markus Schlosser, School of Philosophy, University of Dublin

Discussions of agency tend to go hand in hand with what is known as the causal theory of action. This theory, put roughly, explains agency in terms of causation by mental states, such as desires, beliefs and intentions. An example can be found in the turning of a light switch. I want to adjust the light in this room. That's my desire. I have a belief about how to do that… and then form the intention to do so and then that's what I do.

If the causal theory of action is correct, then it is likely the prospects for artificial agency are fairly good. I think the question then is merely whether we can account for our desires, beliefs and intentions in terms of computation.

But, like so many areas in philosophy, the theory is not without its tensions. Frequent criticisms argue that the theory "leaves out the agent", reducing a situation to a nexus of causal pushes and pulls. Such pushes and pulls leave little room for an agent. It's a reductive theory. If you look at what it says, it doesn't make reference to an agent. It doesn't tell us what the agent is doing, but only what desires, beliefs and intentions are causing. That's how we explain action.

But, what if agency doesn't actually require an agent or a self? While such a thought may seem counterintuitive, agency in spite of a self is actually common thread through many Eastern philosophies. Both Hinduism and Buddhism, along with other Eastern modes of thought, have this in common. It strikes us as extremely counter intuitive that there could be agency without a self, but it's actually a theme that you find in the eastern traditions and where they seemingly agree.

What's more, it really hasn't been explored within this debate. Thinking occurs and acting occurs but there is no underlying thinker or agents. This, too, has an impact on how we view consciousness. According to this view the very idea of a conscious mental state is really a misconception.

Schlosser ends raising the key role consciousness plays in the collapse of the quantum wave function suggesting that we might "really be missing something" given the complexity of the problem at hand.

Concluding Comments, Discussion and Recommendations from Participants

- Steve Torrance notes the similarities between the talk and the thought of Spinoza; a "western philosopher for looking at the basic theory and I think it might take you a long way." He also notes that it is not clear if a machine observing a collapsing wave function would have the same effect as if a human did. Schlosser responds that if one were to look to Hinduism, then the collapsing wave function process would have to be rethought because there are no centres of consciousness.

- Another participant questions the point in "getting rid of the self. It seems like we've lost a lot of explanatory power and we can attribute agency to different bodies in the brain."

- David Miller of the University of South Wales wonders whether "there are other avenues we could take" that do not "dissolve people", such as considering our agency through our relationships with, say, prostheses.

- Schlosser notes that in Hinduism, they do go as far as to deny agency, but says it does come down to how you define things. If agency requires an agent then you must reach that conclusion. "But maybe there's a weaker notion of agency we can work with."

- Jordan Burgess, a research scientist at Amazon, asks whether identifying with the self and a feeling of free will is useful and why. "It's interesting in whether it seems useful or… necessary," Schlosser responds. But, to eastern perspectives, often you don't fully see what the self is, once it's gone. "Only when it's gone do we realize it wasn't necessary and maybe wasn't even useful."

2) Agency, memory and the cognitive ecology of AI. Dr Robert Clowes, Universidade Nova de Lisboa.

Rather than focus on the singularity, in the final session for the first day Robert Clowes from the New University in Lisbon focuses the majority of his talk to "lead you through the AI terrain" as it is today, and examine how this relates to our agency and memory.

Commonly, Clowes says, people attack the ubiquity and constant connectivity of tech today, claiming it to be detrimental to our memory, agency and makes us, overall, shallower individuals. It's a common strand of thought, a response to the fact that "everywhere (in our lives) AI is embedded… so many machines are parts of our everyday life." In many areas we may not be aware of just how connected we are, and in other areas you witness people going to drastic means, such as implanting technology, to achieve this.

Those critical of connectivity often point to what they call the "Google Effect" to show the ills of technology, referring to an experiment whereby people told to memorise lists perform worse when informed they might have access to them at a later date. It has been common throughout history for people to outsource certain cognitive processes and is not, contrary to what many say, a uniquely technological problem. He points to the stereotypical old, married couple whereby one remembers birthdays and the other directions to various places.

Nor is it certain that such connectivity represents an affront to our agency. He points to examples like wearable technology and fitness apps. The recording it does enables a certain degree of reflection on our behaviours not previously possible, and it might remind us that we haven't completed a certain fitness goal that day. A better way of viewing this is that "maybe it is not the case that we're kind of bleeding out our intelligence or our memory abilities into technologies. Maybe they should count as part of us." An exo-self, if you will, one which enhances our agency rather than diminishes it.

However, we do have reason to worry when we do not have access to the technologies that influence our lives. Consider the algorithms that drive what we see on social media that have been making headlines in the past few months. "The danger is that rather than creating a new kind of possibility for reflection, it creates a new kind of possibility for darkness… for obfuscation."

Concluding Comments, Discussion and Recommendations from Participants

- A speaker, who does not identify himself, suggests the talk connects to the interaction between people and prostheses and practices. He says he would add, the paramaterization and externalization of memory is something people have been talking about for a while. It is important to be thinking about the black boxing and hidden aspects of power that are going on within this ecology of technology. "We need to be aware of that and often things that go ping are so delightful that they are transparent in the sense that we don't understand their implications within systems of power."

Identity - Enactivism - Embodiment

The theme of embodiment was raised during a number of the talks and discussion sessions on day one. The first session of day two seeks to examine this particular topic in more detail.

1) Embodied perception and action - a challenge for AI? Dr Simone Schnall, Reader in Experimental Social Psychology and Director of the Cambridge Body, Mind and Behaviour Laboratory, University of Cambridge

Building and expanding upon a theme encountered in earlier talks, examines the role of embodiment in cognition, in particular looking at perception and action in relation to this. Embodiment is, to Schnall, crucial in understanding cognition. "One has to look at the body to fully understand cognition," she says.

In a cognitive sense, earlier works suggests that "one cannot separate perception and action". In particular, affordances, the possibilities of action, can have great influences on our perception. Consider the legal case in Germany where an artist was sued when members of the public kicked concrete balls placed on the floor; people were unable to resist kicking what is, in reality, just a sphere, and naturally broke feet and ankles. "This is a really good example of an affordance that is evoked by an object… a sphere can become a ball when it has an action affordance." In short: context matters. "It all depends on how one can act in that context".

In more natural settings, Schnall and colleagues have shown how perception can alter based on context and what is going on in the body. Participants in an experiment, when asked to comment on the steepness of a hill, responded differently when they had lower blood sugar or carried heavier bags, for instance. Social context also matters; when social help is available the hill appears less steep. "Our interpretation for that is that there is a really strong connection between perception and action and cognition to how people think about the world, how they act in the world and how they perceive various obstacles and challenges".

Even for very high-level cognitive processes such as language, we still tap into the sensory or bodily systems that will be linked with said actions. "Word and action is closely linked as far as the brain is concerned," says Schnall. Such findings have been applied to try to help those unable to talk (such as those in vegetative states) communicate by getting them to think about certain actions or

behaviours. For instance, one might teach a patient in a vegetative state to use the mental image of playing tennis to mean 'yes'. "Again, it's very difficult or almost impossible to separate out the perception and the cognition in the action."

This all has implications for how we think about AI. We can't just think of the brain or as intelligence as the brain in movement because we must also think about the body and about the affordances for action. "Do you figure out the chess playing part and then tack on the body in some way," she asks. "That is not how biological agents work. So what does one do? This is of course the big challenge: how to actually apply all this in the context of artificial intelligence."

Concluding Comments, Discussion and Recommendations from Participants

- Anders Sandberg wonders when we have tools that extend ourselves in some way, what happens when the tools have a bit of autonomy? For instance, when you are driving a car your sensorium does, to an extent, extend along the car. What is known about this? Schnall thinks the work is very important area of research, pointing to studies where a baton might be included in a body's schema. "Is it just physical objects," Schnall asks, "or could this go to cognitive skills too" Consider outsourcing memory capacity to Google, for instance."

2) RECtifying intelligence: radical enactive/embodied approach to cognition and intelligence. Prof Erik Myin, Professor of Philosophy, Director, Centre for Philosophical Psychology, University of Antwerp

REC stands for the Radical Enactive (or Embodied) approach to Cognition and intelligence. The "main claim is that the vast sea… of cognition is contentless. But some cognition is contentful." Cognition, explains Erik Myin, can be thought of as "the intelligence interaction with the environment," and thus includes actions such as reaching for a cup, an animal catching prey and solving an equation on a blackboard. "Cognition is in the interaction," Myin says.

In addition, "all intelligent interaction with the environment is intentionally direct." It targets specific things in the environment.

And for cognition with content, this "consists of specifying, portraying, describing something in reality or not in reality… it's not just a reflection of reality, it's not just a correlation with reality but it's something that can have saying power of itself." For instance, one could say it is snowing in Cambridge today, and there you are portraying the world with specific content that could be a description of the world, but could also not be.

So what is perception and what is perception using this view if it is not conveying contentful information. A way to think of it is to think of it as a reaction.

So how do we explain content? We can explain content in language but not perception. In language you can account for it, but you cannot do so for perception without reference to some hitherto unexplained primitive form of representation. You don't have that story for content in

perception.

Ultimately, then, cognition, including perception, is intentional without involving contentful representation or computation. This does not deny the fact that computational cognition or content involving might exist, but in order to do so it must be supported by some faculty such as language. "Computing is like representation, it's not a given thing found in nature, neither in rivers, in planets nor in brains. It's not even in computers."

Concluding Comments, Discussion and Recommendations from Participants

- A commenter, says a general comment for the whole workshop could be that "we should be making a clearer distinction between the possibility of artificial general intelligence and artificial intelligence that's actually being built today." Would going into rooms of people developing, say, driverless cars and saying they need to overhaul their language actually be useful?

- Marta Halina is "very much in favour of embodied cognition," but admits she's never really "been on board with antirepresentationalism." "I have trouble seeing how it can account for cognition and bodies in the world." Myin responds that he is not denying those models, just "a certain way of seeing them."

- Andrew Brown, a science journalist, notes a similarity between Myin's theory and the Eastern philosophies we have encountered earlier in the conference.

- Ron Chrisley feels the burden of proof for non-representationalist views has not been met, whilst it has for those thinking representationally. "They've shown its utility; the challenge to the non-representational is to say why it's not valid".

- Rob Clowes believes that the endeavour Myin describes to be of dubious scientific merit: "from a philosophy of science standpoint I don't think anybody doubts you can tell different stories about processes… merely saying that we can offer another explanation of the same thing without using the word content is not really a scientific programme as such… if it's merely that you're trying to get rid of certain vocabulary, it's not clear what the scientific contribution of that is."

- Sally Davies, from Aeon magazine, agrees with Clowes. She has sympathies for what motivates the non-representational agenda, "but I still really struggle to understand what's at stake in saying cognition is contentless as opposed to redefining or reconceiving what we mean by representations and content… the definition you've provided risks being a straw man."

Identity and Self

1) Someone but not a Self. Dr Daniel De Haan, Research Fellow, Ian Ramsey Centre for Science and Religion, University of Oxford

Frequently, we find ourselves asked questions pertaining to our identity. With similar degrees of frequency we inquire into the identity of others. Questions such as "Who are you?", "Who am I?"

and "Who are we?" and their answers form the subject of this talk.

There are three ways of answering these kinds of questions. We may answer existentially, questioning how we see ourselves in reality itself, what values we might have and what it is to live or die well. We may answer practically, such as the responses we give when others ask. "What do you say when you introduce yourself to others?". Finally, we may also answer questions theoretically. It is this last way, the theoretical, that forms the subject of this talk.

One might answer this question in one of two ways. There is the physical sense, placing and identifying oneself according to science. There is also a more "common-sense" answer, that more anthropological approach where we are "situated somewhere within that kind of (social) context". However, it's important to note that there are many different approaches to be found within these broad categories. We make things too simplistic and easy for ourselves when we presume there is such a thing as a standoff between the scientific image and the common sense. And, De Hann elaborates, there are all sorts of distorting factors occurring within such enquiries that distort our considerations.

Ultimately, we often find ourselves answering the question "who we are?" with reference to a self or a someone. "Why do I start with the self or start with a someone", De Haan queries. We often use such terms interchangeably, but to De Haan there are actually important distinctions. Importantly, De Haan claims that in order to think of ourselves as a self we must have already considered ourselves as someone, an individual in constant development throughout life. To answer this question actually requires adopting the stance of a someone to understand why you even adopt the stance of self. You need to think of a story about your own life and your own sort of inquiries and how they lead you to beginning to think about yourself as a self as opposed to a someone.

In his concluding thoughts, De Haan relates this to AI. He notes that while not all AI researchers would be interested in such questions, those who seek to model or emulate or simulate humans would be interested in examining their conception of human intelligence. "If they are interested in that, is their conception of human intelligence the intelligence of the self or the intelligence of a someone and why?"

2) Corporate individualism, robots and AI: a call to Martin Buber's I and Thou. Prof Kathleen Richardson, Professor of Ethics and Culture of Robotics and AI, De Montfort University

René Descartes' famous *cogito* is perhaps the centre of ego-centric philosophies; we, the self, can and do exist in isolation and without reference to another. We might contrast this ego centrism with a more interpersonal philosophy where we do not exist in isolation from others. The thought of Martin Buber, to whom the word I represents a combined word of I-it and I-thou, is a good example of this. "We do not exist alone and we are always with another and how we characterize the other is profoundly important."

For me the idea is that we can't ever speak about ourselves as alone. So if we are invoking another, it's important that we know who we're invoking and what we are actually invoking. We

might look to how people are connected on social media and these connections are representations of what people think relationships between humans are really alike. Importantly, Buber differentiates relationships between other people and relationships between things; he rejected the idea that people are property, that people are things.

This has important consequences when thinking about AI and robotics. We have a problem on our hands because apparently one of the rationales for developing a culture of robots and AI is the looming crisis we are entering: a demographic time bomb, that we're running out of humans. That's the narrative, anyway, created despite the seven and a half billion humans on the planet. "The solution presented for this narrative, that we must develop robots and AI, particularly around care around this intimacy and interpersonal relationships." Richardson says she can't help but notice the irony of "how many robots in labs look like children" created by many scientists who "don't actually do any priority child care. But they love to come into the lab and talk about the robot children."

Ultimately we actually only survive because of our relationships with other people around us, and we're thinking we can replace other human relationships, intimate relationships, with artefacts, machines and AI. Examples of tales where children have been raised by animals or in states of deprivation, as well as animal studies doing the same, highlight that "it actually wasn't enough to just have a substitute… there needed to be something more."

"If we do believe that human beings are interchangeable with other entities then I think we're creating a society of problems for ourselves."

3) Robots without families: on identity and organic continuity. Fr Ezra Sullivan OP, Faculty of Theology, Angelicum, Ponitifical University of St Thomas Aquinas, Rome

In thinking of the singularity we might consider an AI of equal or greater intelligence than that of ourselves. Rather than simply being able to undertake various computations or pass various tests, such an AI would require some kind of interiority. It needs to have something of what people call sentience in interior life and personal identity. And so the question becomes, is that kind of artificial intelligence possible?"

A key part of a human's interiority and identity comes from our nature as organic beings. We identify with our more immediate family members, but also with all other living beings through this shared organic embodied nature. We belong to what one can call the family of the living. We have organic continuity.

We might try to rebel against this, embrace it or exhibit indifference, but it's impossible to truly remain unaffected. In the end we are still affected by our biology. And, it follows, things that differ biologically think about themselves differently. This process of self-reflection one recognizes one's genetic inheritance may be slightly different than another's and that entails differences in one's practical daily life."

Put briefly, our nature as organic entities, and our connections and differences in respect to all other organic entities, plays an important role in our conceptualising identity. Robots, however, do not have this, and this could lead to important differences in identity. They don't belong to any

organic family. They're constructed. They do not share the common makeup of cells, they may receive transplants of limbs or other parts without fear of rejection, and their software might be copied endlessly into different vessels. While we inherit our genetic information, the information making up an AI is given to it by a human. AI and robots do not develop, grow and change over time in the manner of humans and other living things. Many questions that are perennial amongst humans, such as the existence of a soul, would, Sullivan says, simply not apply to robots. And this entails important differences, potentially impacting how robots and AI might behave.

So, Sullivan reiterates, there exists this gap between humans, and other living things, and the artificial. The organic and the inorganic. How do we bridge this gap? We might institute a developing and nurturing period for robots. If they don't grow up like little children perhaps we can treat them like children for a little while. We could change the software, possibly giving it some kind of false history. We could, as in some science-fiction stories, develop a biological basis for robotics. However, it's unlikely these approaches would be successful. AI would, given their intellect, be unlikely to see us as equals and would soon see performance differences. But I think it's much more likely that robots actually would see themselves as above us. They would be stronger, of smarter self-control, more powerful; they would know that they're not like us. They would use that knowledge to their advantage.

Robots differ from humans greatly because humans are animals and robots aren't and things that differ biologically think about themselves differently and they act accordingly. Therefore robots will always necessarily act and think differently from humans. And I contend in irreducibly important ways."

Concluding Comments, Discussion and Recommendations from Participants

- Anders Sandberg, of Oxford's Future of Humanity Institute, says he agrees with the syllogism, but notes that there are many positions aside from equals, betters or lessers. "There are lots of webs of connection," he notes. "What about robots beside us or in the fourth dimension?" he asks. "There are lots of ways of being different," he notes, and suggests that in a familial sense computers might have an analogical family or relatedness to, say, the programming code it uses. "It can be very plausible that the robot will not feel that much about what language it is implemented in. But we can't really rule it out."

- Geoff Carr, the Science Editor at the Economist, says he found the talk interesting from a biological point of view but that, from this point of view it was "all looking backwards." Reproduction is crucial in biology, "but if you have an intelligence embedded in something that's not reproducing its intelligence won't have been constructed in the context of having to reproduce." He finds this distinction more important than those made in the talk. Much of what we describe as bad in human behaviour often comes down to acquiring reproductive advantage. Sullivan responds that this does raise the question of so called "free agents", for instance celibates, and whether they might be freed up from the ties many other biological creatures worry about.

- Fenella Cannell, a Professor of Anthropology at the LSE, noted the synergy between the three talks, but would like to pick up on a theme in the first. She says anthropology particularly critiques overly narrow descriptions of what it means to be human, narrow definitions that today have over-

elaborated on considering the self as an isolate. "This can skew and limit what we understand humanity is. The fantasy of the asocial person is a particularly dominant and tempting error of the contemporary West." We must be careful not to project this model of humanity when engaging with the development of AI. De Haan says that it (anthropology) is a "very helpful addition" but notes that the discipline is neither widely read by those outside it and that it is vitally important for anthropologists, with that in mind, to engage in meetings like that of today. Moreover, beyond attending interdisciplinary meetings, De Haan says anthropologists and other academic disciplines need to make efforts to publish outside their usual venues so as to not get 'siloed'.

Richardson responds, believing it an "absolute fantasy to think a commercially produced artefact that's been developed under capitalism is somehow going to have consciousness and fly off and develop its own way of life." It's an area, she stresses, that really needs addressing, and one that certainly needs other disciplines to participate more. "And particularly women," she adds. Having been developed out of finance and military systems, she explains, what we're "really developing is corporate anthropomorphism--commodity fetishism—by another name."

- Simon Colton, of Queen Mary University in London, wanted to speak about the notion of presence. He and colleagues have been trying to develop an AI to have presence in a creative community. He wondered if could expand on this with respect to the self, noting being in a room with a screen that has a presence and a turned off robot that clearly does not. It may be far from personhood, he says, "but could it have a presence and what would it have to do… to convince people of that?" De Haan believes Colton already described ways in which you might have presence, referring to the affordances with robots that do and don't work. There are all sorts of questions, though, like whether it has to be living or animated, but there will likely be metaphysical disagreements. He notes contrasting affordances would be a good place to start.

- Richard Watson, with Tech Foresight at Imperial, wonders if there's a piece missing. He says this very much seems to be a discussion of them and us, and wonders whether it is impossible to "flip this" to a scenario where we add machinery to ourselves and become more machine like. "What are the ethics surrounding that? Where would that fit?" he asks.

- Antonio Ramos Diaz points to cases in, for instance, Japan whereby individuals engaged in what are objectively I-it relationships but they believe they are engaged in I-thou relationships. Particularly, he points to cases where people are in relationships with virtual beings via apps. Richardson says the question is really complex, and time limitations prevent a reply. She does, however, note she has a book out on sex robots coming out next year that will address this point.

Ethics

1) Brain-Computer interfaces: new varieties of cognitive imbalance. Dr Karina Vold, Faculty of Philosophy; Leverhulme Centre for the Future of Intelligence, University of Cambridge

As many talks in this series have illustrated, many are now pushing back against the traditional Cartesian image of the mind situated in the brain, instead opting for more embodied, situated and enactive models of cognition. Thinking in a more situated sense, it is possible to identify ways in which technology may contribute to our cognition; this could be simple technologies such as a pen

and paper, which, according to Karina Vold "can serve some of the same functions as our brains " or more advanced cases of brain computer interfaces that we are starting to see today. These brain computer interfaces, or BCIs, are the subject of the talk.

Put roughly, there are three types of devices that fit into this camp of brain interfaces. There are output devices which are aimed at influencing brainwaves. There are input devices, where brain patterns are detected, interpreted and recorded. This is where a lot of interesting work is going on today. Finally, there are bidirectional devices, which we can call the "gold standard." These can both send and receive and are, of course, the most difficult.

Within these categories we might also identify two types of device: invasive and non-invasive. The former involves implantation into the brain, and obviously entails many risks to the user. The latter does not actually interfere with the brain physically and might, for instance, constitute a magnet or a monitoring piece of headgear.

There are many kinds of applications. Rehabilitation to help people is a common and sought after application, such as helping to control prostheses or wheelchairs. Others include assessment and revealing purposes, such as assessing consciousness or levels of wakefulness. The application to focus on here is enhancement. What's the target of enhancement? The short answer is basically every sort of cognitive trait you can think of is being targeted in one way or another for enhancement. For instance, BCIs can be used to enhance memory, or they might be used to add some new kind of sensory modality, like infrared sensing, that we do not currently possess.

Naturally, enhancement raises a whole host of ethical questions. Pre-enhancement we have issues of informed consent and of the risks implantation might present, as well as pressures people might feel to be enhanced. But post-enhancement is also an ethical minefield. What pressures might an enhanced individual face if starting a new job, for instance, whereby they might feel or be obligated to remove an enhancement that helped them get the job in the first place? There are also serious privacy concerns. This very personal data could tell a whole story about you. So who's storing it? Where is it being stored? Who has access to it?

There are also very real risks about brain "hacking" and "jacking". A lot of these devices are communicating information wirelessly and what we know is that any information that's communicated wirelessly is susceptible to hacking. There's also the risk someone might be able to gain control of your mind through similar means, just as one does today with a computer. This could even be subtle. If an action can be influenced in various ways I think we have to question whether it is an autonomous action.

There is also the risk of what Vold calls cognitive imbalances. We might throw off the natural balance of the human mind. For instance, how will our decision making differ now we have so much more information with which to make it? 360 degree spatial awareness could lead to the diminishing of other, such as somatic, awareness.

Vold takes pains to distinguish her argument from what she terms the argument of hyper agency typically encountered. Such arguments warn we should not interfere with the human mind as a matter of principle, that we would be upsetting something natural and that should be appreciated as

such. I'm not against enhancement in principle, Vold says, "but rather as it's touted as enhancement and ends up being a kind of diminishment."

Concluding Comments, Discussion and Recommendations from Participants

- Iason Gabriel of DeepMind wanted to know more about using BCIs in the context of moral reasoning. Vold points to experiments using transcranial magnetic stimulation that have shown the effects of this on moral reasoning by suppressing certain beliefs.

- Keith Mansfield, a writer and publisher, wondered in terms of cognitive imbalances is there any idea on the imbalance between those who are enhanced and those who aren't and what that would mean for society? Vold recognizes that there are genuine issues that would arise between "haves and have nots". "There's clear questions of disparity here."

- David Mellor, from the University of South Wales, wants to pick up on the socioethical aspects of the talk. He is particularly interested in the ethics of invention and building in responsible research and innovation. Vold says there's "a natural reason to think this will be a problem… it's a lot easier to release something on consumer markets" than to approve something for medical or rehabilitative use.

- Sumit Paul-Choudhury, the former editor in chief of New Scientist, asked Vold to comment on the similarities on the ethical side of chemical cognitive influencers. "There are some striking parallels and striking differences," he explains. There are, Vold says, overlapping sets of concerns with other enhancements, but not identical ones. The bidirectional communication possible is the key difference here. Nootropics, for instance, don't have the possibility of feedback loops, whereas this technology could offer kinds of feedback loops.

- Sally Davies, from Aeon, finds herself intrigued by the incommensurability of values given that your standards of determining good will differ depending which side of the enhancement gap you are in. How would you make a decision in that situation? Vold, in her response, likens the scenario to pregnancy, whereby becoming a parent causes a total shift in your values. "You can't make a rational choice about it… you can't know all the information that's relevant in what's typically thought to be a rational way."

- Michael Stevenson, from OECD Education, is worried that "enhancement threatens the equity principle in education".

2) Ethics for E-Persons. Prof Steve Torrance, Visiting Senior Research Fellow, COGS, University of Sussex; Professor Emeritus of Cognitive Science, Middlesex University

In recent years, recommendations have been put forward to the EU Council raising the possibility of awarding some kind of personhood to AIs or robotics, an e-personhood. A "proposal contained this possibility of paving the way to a form of legal personhood for robots". Such a proposal is, Torrance explains, controversial, but also something that could be achieved with relative ease within

just a few years. It is therefore worth dedicating some attention to the ethical and social issues e-persons might raise.

Typically, our notions of moral status are rooted in biology, and, in particular, consciousness and sentience. We suggest people have certain moral agency, in having the responsibility to act morally. We also have what we call moral recipiency, whereby one has certain rights afforded to them, such as freedom and autonomy. Being non-biological in nature, even leaving aside questions of sentience, it is difficult to ascertain the legal and moral status of machines, given how much of our understandings of this are rooted in biology.

Regarding legal personhood, we currently use what they call a natural person model. Legal status for a robot cannot be derived (from this) since the robot would then hold human rights such as the right to dignity, the right to its integrity, the right to remuneration, or the right to be a citizen, thus directly confronting human rights. This would be in contradiction with the charter for the fundamental human rights of European Union. But, Torrance questions, the potential for some kind of e-personhood does raise the possibility of a secondary moral status.

Consider a legal scenario whereby a machine is afforded legal ownership, say over the proceeds from trading on the stock exchange or producing a painting. Would it be morally wrong to steal from such a machine? Does the mere fact that the robot is a legal owner of certain property mean that if I were to go into the robot's house and just pick up the laptop from the table and take it out and just arbitrarily expropriate it make it morally wrong? Would this have wronged the machine or just suggest something about the thief's moral character? When you think of it, the legal right to own property of the machine may generate some secondary kind of moral status, a right not to have its possessions arbitrarily taken. It may also generate other moral rights and duties, such as the duty to care for a pet it owns, for instance.

There's a kind of secondary moral status which is not anchored in consciousness, but is rather more anchored in the legal status of being a property owner. Such a scenario is one we could imagine occurring with relative ease and in not a great deal of time. Unlike many of the other developments that we looked at, it would be relatively easy for a legislature or two to say 'well you know this robot has or this system has traded on the Stock Exchange and has earned a lot of money for its owner and should be allowed to act to be the possessor of this wealth. It's actually conceivable it could happen within a few years let alone decades. They may actually amass enormous amounts of wealth and might actually take over the financial system.

Concluding Comments, Discussion and Recommendations from Participants

- Simon Colton, who has an AI that makes art, acknowledges he is in the moral quandary posed by Torrance's talk. "My guess will be that morality will emerge from practical matters." He notes that his art is worth more when owned by the software. "No one wants a picture from a computer scientist like me but they might want a picture from a piece of software or a robot." Another issue is that code needs maintenance. One might put money into a pot to maintain the code. It is, Colton says, morally interesting if someone steals from that pot.

- Kathleen Richardson cautions against the use of science fiction in these discussions. "These

things aren't transferable." Richardson also points to the new kinds of capital that have emerged and been cultivated, such as the ability to purchase 'plots' of cyberspace. In response to the science fiction point, Torrance notes how influential it is, even in the European recommendations one sees overt references to sci-fi. "It's clearly very potent and can often play an exaggerated role in our thinking, but also it gives us thought experiments that, if we're careful, can extract important ways of thinking about what actually can happen in the near future."

- For Keith Mansfield "this seems pure simple discrimination against non organics."

- Margaret Boden outlines that in Japan their attitudes to robots differ greatly to us. A number of robots there have been given various sorts of legal status, such as being made legally part of a family group, which is especially important there and some sort of license to reside in a locale. "What I don't know and would really like to know is if there have been any legal arguments coming out as a result of those decisions and, if so, how they've been handled and who has won."

- The final question asked, from a person who did not identify themselves, suggests that rather than belonging to any individual, perhaps it is society as a whole that ought be the recipients of the benefits that come from a robot's labour.

Aesthetics and Imagination

1) Should an AI cause insult and harm? Prof Simon Colton, Chair in Computational Creativity, Queen Mary University of London

Today, artificial intelligence is the driving force behind many products considered creative, such as art, works of literature and new products. But what would the ramifications of truly creative software be? Undoubtedly systems today can cause harm, Colton says, pointing to the humiliation AlphaGo might have caused Lee Sedol and the potential for killing individuals that driverless cars possess. Of course one could argue that they are doing more good than harm which is undoubtedly the case. But it's also true that their systems will cause harm. So are our tech leaders (who frequently preach for AI systems to 'do no harm') saying that AI systems causing harm is a good thing sometimes?

Moving on from death, injury and humiliation of the previous examples, what about an AI merely causing offence? The answer is, "Yes. Of course they should." Causing harm is sometimes a necessary evil and causing offence is definitely a desirable outcome.

Colton points to a Twitter bot that produces posts on fictional shooting responses by the NRA. Obviously, the NRA is less than pleased and have taken offence to such a bot. They are offended, naturally, but others also view this as productive activism. Being offended is very context specific and subject specific. Colton also points to the initial reactions Parisian audiences had to the first Impressionist paintings displayed there. Audiences were so offended they threw shoes and canes. Today we view these works as treasures. It goes to show that we don't always know what we will value in the future, it could even be something we hate and revile today.

To build only AI systems that do not offend unnecessarily restricts the good we might derive

from them. "Building pink and fluffy AI systems to be nice all the time will severely limit their power and value within society."

But lacking authentic life experiences hinders the ability of AI to be truly creative and offensive. He illustrates with reference to racist or homophobic comments. Such comments lose their bite when one discovers they were written by an AI system rather than another human holding those views - it's inauthentic. This issue will remain even as AI systems grow more advanced. AI software doesn't live in the world like people do and even when robots roam the streets they will still be very different entities to us.

Such inauthenticity is a major hurdle for computational creativity systems. If written without the backing of authentic life experiences, then products lack authenticity, and will not truly be able to offend us. We can't truly be offended by something if it was written by clever natural language processing algorithms.

To be more authentic, "creative software needs to have and record authentic life experiences… albeit ones which differ from those people have". Regarding offence, in order to be considered truly abusive or offensive "it will likely need to have recorded life experiences in human society on which to base abhorrent views… it needs to be affected by the things that it writes about and have a model of human emotion."

Ultimately, Colton hopes he has made the case for "investigating how to implement dodgy software behaviour with applications from saving life to making people laugh". If we only produce that which does not cause offence we are "severely limiting" the benefits we might derive from AI systems. We "will miss opportunities that will enhance many more human lives than it would degrade."

Concluding Comments, Discussion and Recommendations from Participants

- I'm worried that you're turning into Jeremy Bentham, says the first speaker who did not give a name. I've never liked utilitarianism as an attitude for how society should change… there seems to be no persuasive case for why it would work now. It matters to an individual if put out of a job - they only have one life, she explains. "I felt that your presentation… slid over these questions. The fact that human suffering is part of human sociality."

- William Clocksin was "puzzled by why you were interested in developing machines that could offend people."

2) Imagination and Memory. John Cornwell, Director, Science and Human Dimension Project, Jesus College, Cambridge.

Journalists often ask how people feel about something, and I'm going to pose that question about AI, Concern surrounding AI is not just limited to university seminar rooms but does, in fact, occupy the minds of the general public too. Personally, Cornwell feels a "mixture of discomfort, foreboding, melancholy, existentialist angst... Something to do with the autonomy and responsibility

challenged." To a degree a great deal of this is still in the imagination as it hasn't happened yet. Parallels might be drawn between attempts to create AGI and Mary Shelley's Frankenstein.

In humans, memory and imagination are closely and intimately linked. Alzheimer's, for instance, not only negatively impacts memory formation and recollection but also future planning and imagination. There are also shared neuronal correlates between memory, future planning and imagination. It's illustrative of the reconstructive dynamic nature of the mental activities involved, its dynamic constructions and reconstructions. In what sense is this imagination? Could this be a property manifested in AI systems like AlphaGo? The imagination ceases to function as a mirror reflecting some external reality and becomes a lamp which projects its own internally generated light on things. Now was this productive creative imagination a property manifested by AlphaGo? Has it gone, with intuitive and imaginative play, beyond some concept of objective merit and been elevated to a position otherwise occupied by humans?

Professor Margaret Boden, in an earlier talk, offered us some consolation. She said that "while in human affairs the issue of something or someone mattering is of significance, this could never apply in the workings of AI systems for which nothing matters." One might feel anxiety or stress when presented with urgent tasks of import, and while an AI could prioritize tasks it would not be genuine anxiety. "The computer couldn't care less". Similarly with betrayal, says Cornwell. "It's an emotionally imagined experience that would elude an AI system."

Cornwell notes, through correspondence with Rodney Brooks and John Naughton, that while Lee Sedol, in the famous match against AlphaGo, received support from a "cup of coffee", his opponent received help from 200 people. "That's not superintelligent. That's a super basket case." AlphaGo had no awareness of the world around it, or even that it was playing a game that took place on a 2D board. "They operate as transactional programs that people run when they want something."

Ultimately, is the power of imagination to free and bless humanity at risk of the machines we are creating, that which cannot genuinely replicate parts of the human imagination? Does that final power to free and bless humankind stand in danger of being rivalled by self learning machines? The answer is, it depends.

Facing the era not long in coming when humans will share the planet with alien intelligences and imaginations and intuition in some way is superior to our own, suggests an opportunity to explore and interrogate ourselves yet again on the nature of our own imaginations and intuition.

A theory of imagination today would draw on a variety of disciplines. Should we be comforted by the reflection that our imaginations will always outstrip that of the machines on the level of addressing the meaning of life? Or will we continue to outsource our cognitive and moral baggage to machines that lack a sense of the meaning of life, and are an intuition that is as uncomprehending as well as incomprehensible?

So what do I feel? I feel a sense of increasing melancholy and angst along with the optimism.

Concluding Comments, Discussion and Recommendations from Participants

- "I think (AGI) is a chimera… a fatuous project," says the first commenter who did not state his name. "But I am at least as melancholy as you." He goes on to say that today's technology is leading to a "programmed world which will ultimately be populated by predictable and programmable people." Concerning ourselves with the future is a fatuous project, he explains, when it is really "distracting our attention to what the thing is doing now."

- Simon Colton: "as a scientist I want to apologise for the fact that as scientists we've let the impression the rest of the world has about AI get out of hand. There are reasons for this," he says, noting that in order to get sales businesses need to oversell capabilities, in order to sell papers journalists need to exaggerate abilities. "But we as scientists should be out there more explaining what actually is the state of the art and that it's not really at a stage where we should be concerned."

- One commentator, who did not give his name, feels that a lot of the more pessimistic predictions "cannot be based on scientific thought… neither is it philosophical.' He wonders whether it "comes down to a religious point of view."

- From a caring perspective, AI enables us to place people in communities where we otherwise wouldn't have been able to, says a woman who did not give her name before speaking. "We're excited by enhancing humans to overcome disadvantage," she says. It can enable people to have independence. "What we are frightened of is enhancing human capacity to create super races."

- Bonnie Zahl of the Templeton World Charity Foundation says that she cannot help but think about children. "My first feeling is that I worry for my children. I have no ideas what their lives might look like," she explains. "If there's one thing I can do better than a machine it's that I can be a better parent than a machine," she says.

- Simone Schnall says the question of 'what are we doing with AI' has really resonated with her. Are we modelling human intelligence or are we just building systems that can solve problems? It's almost impossible to not take the human being as the default, she says.

- Ezra Sullivan says that we haven't really discussed that AI will, generally, offer us more leisure time. To Aristotle leisure is valuable giving us time for art or philosophy, but today we end up looking at videos of dogs. "What are the consequences of this?" When listening to Colton's talk , Sullivan says it struck him that why teach an AI to paint - why not teach the people to paint. "I think one of the biggest issues is that we're commodifying ourselves."

- Beth Singler, at the University of Cambridge, says that many questions end up asking what is a human as well as what is a human for. It's been answered in numerous different ways over time and space, "but we've started to move into a void space where we are less certain than before. It's easy for some groups to fill that." Going forward, Singler says, we need to "reidentify" that direction of what we want humans to be.

- Alicia Perez, a Chaplain at the University of Manchester, says she has an optimistic view about

how "artificial intelligence might be a natural evolution of being human." She points to chess players capable of improving their game through learning from machines. This is the positive aspect. The question is what will we do with our time?

- Michael Stevenson, from the point of view of education, is "feeling challenged." He says that policymakers around the world are doing the right thing in addressing what "children should learn to nurture what's distinctly human…. But there's a long way to go."

- Antonio Ramos Diaz has "mixed feelings". He notes that he rarely experiences boredom anymore, which would have usually inspired him to create, due to the ever present flow of information at hand. It's led to a distracted and "less deep" culture.

- Marta Halina wants to provide a counterpoint to the earlier confidence that AI will not have human level cognition in any significant sense. She notes that the difficulty of tasks, such as drinking a cup of coffee, is dependent on whether you're looking at it from a human or an AI perspective. We might, Halina quips, "say that humans are just not well designed to play Go," pointing to the fact that many of the best players train so intensely for this they often forgo other education. From an animal cognition perspective, we are frequently surprised with abilities. If we are not to worry about the risk of AGI, we should do it for reasons other than confidence it will not happen.

- Julian Hughes, a psychiatrist, says that from a philosophical point of view, he sees radical differences between machines and humans. Looking at worries, he says he just doesn't see why people are worried about a computer learning to play games like Go. One really beneficial use of AI, Hughes explains from his professional experiences, is to try and alleviate the loneliness many people face. "But it would be better if we didn't have to rely on AI."

- Yaqub Chaudhary, from the Cambridge Muslim College, thinks the issue of child development is particularly important and we need to think how we treat and interact with these artefacts. There is also a lot to be learnt from history in both ourselves and this technology.

Project overview

Artificial Intelligence (AI) has entered a new era in which machines are no longer constrained by programming, but exhibit self-learning capabilities with ever increasing speed and capacity. The aim of Artificial Intelligence at its most ambitious is to achieve a "Singularity", or Artificial General Intelligence – i.e., to outstrip human intelligence. Future goals of AI research speculate about the possibility of a "transhuman" condition—in other words, enhanced humanity. In this project we aim to draw the humanities into the debates and critiques of these ambitions. The disciplines of the humanities envisaged include literary studies, philosophy of mind and of religion, anthropology, cultural studies, and theology.

So far the debates about the impact of AI and critiques of AI's influence have been largely confined within the disciplines of computer science, social sciences and economics, and therefore primarily quantitative in emphasis, as well as being primarily focused on practical applications such as security, defence, medicine and social welfare. This is clearly visible in texts such as Martin Rees's *Our Final Century* (2003), Ray Kurzweil's *The Age of Spiritual Machines* (1999), Nick Bostrom's *Superintelligence – Paths, Dangers, Strategies* (2014), and Murray Shanahan's *Technological Singularity* (2015). There is therefore an urgent need for well-informed qualitative assessments of the potential impact of AI from the humanities, including from theology and spirituality.

Artificial General Intelligence works on the hypothesis that technology will replicate, and even outstrip, not only human intelligence generally, but specific human faculties such as imagination, consciousness, and agency. To what extent will these ambitions match, challenge, demoralise, or perhaps even aid these faculties as understood by disciplines within the humanities, including those that deal with moral and spiritual dimensions of life?

The goal of the AI and the Future of Humanity project is to explore this and related questions through a series of three conferences that bring together representatives of AI research along with scholars from the broad span of the humanities. Our target audiences for both the meetings and the subsequent books include academics, researchers in these fields, religious leaders and the wider public. The latter will be facilitated primarily by the journalism produced by representatives of the quality media who will be invited to the meetings, and by the books published after each conference. Outputs of the project include three major conferences; publishing; conference reports; short films featuring Q&A with speakers; and a range of print and broadcast journalism, including reviews, feature articles, op-ed pieces. We anticipate that the project will strengthen existing and forge new links between the AI communities and the humanities to ensure that the humanities - including philosophy, anthropology, and theology - become part of the conversation and the debate about artificial intelligence and the future of humanity.

Big questions the project is addressing

AI researchers are attempting to devise machines that exhibit not only intelligence, but intuition and imagination, capacities that may lead to analysis, decision-making, and forms of creativity that equal or even outstrip human faculties. Future machine intelligence might challenge, and perhaps threaten, human intelligence, wisdom, and creativity, or it might enhance these faculties. The humanities offer an ideal vantage point to assess and critique what AI research means by intelligence, intuition, imagination, wisdom, and creativity, and as such they can help to clarify the challenges, benefits and threats that machine forms of these capacities represent.

How does *imagination* in AI terms compare and contrast with imagination in the arts and aesthetics? How does *creativity* in AI compare and contrast with creativity, and *creation*, in the arts, sciences, and theology? Is consciousness a crucial dimension of decision-making and creativity? These questions are crucial for understanding in what sense or senses artificially intelligent machines will exhibit moral and even spiritual behaviours, values, and even virtues. A machine that exhibits forms of imagination, creativity, agency, consciousness and autonomy would indicate the emergence of an intelligence capable of being in a moral, cultural and social relationship with humans and, some would argue, in relationship with God.

As well as asking how we expect the machines to behave towards us, we must ask how we would behave towards them? Should we create AIs whose consciousness might well be unbearable to them? What ethical principles should be invoked in the case of terminating a machine that possesses a form of consciousness? Additionally, how might human identity, viewed by some as made in the image of God, be altered in view of our ability to create machines in *our* image? Also important is the tendency, recently noted by Jonathan Sacks, to outsource human identity and delegate decision making to machine systems. The more humans delegate decisions to machines, the less autonomous we may become. How should we use machine decision-making to our advantage and avoid the outsourcing of our moral decision-making?

Statement of significance

This project is about navigating our future in the light of major changes and challenges presented by developments in AI. Much thought and work is being put into the practical implications of AI for health, employment, food security, poverty, education, defence, and climate. Our project is concerned with the impact and scope of future AI on deeper cultural, moral, religious, and spiritual dimensions of life. The project seeks to make connections between AI and the humanities, in part by making the future of AI accessible to wider constituencies that include the broadest span of the humanities, the world's religions, and the many expressions of spiritual and religious life significant for human self-understanding.

The stated goals of Artificial General Intelligence - the "Singularity", or "superintelligence", where machines will achieve "human equivalence" and beyond (e.g., Kurzweil, Bostrom, Shanahan) - imply unprecedented human responsibilities for deciding such issues as the goals, regulation, extent, and limits of future machines. In debating and critiquing the qualitative as well as the quantitative significance of future AI, we are insisting that cultural, philosophical, and religious perspectives not only have a part to play, but bear significant responsibilities, in understanding and hence helping shape those goals and limits.[1] We believe that through encouraging the humanities to play a part in AI
developments, we shall encourage the development of a fresh understanding of what it means to be human (both individually and socially) in the light of machine-human continuities and discontinuities.

Given the deep implications for the human spirit, the present project has relevance and significance for the widest constituencies of people, present and future, old and young. At the outset we aim to involve, and make connections between, the AI community and representatives of the humanities on the other hand. In this way we hope that the two sides will speak to each other, influence each other, and ultimately help each other. At the same time, we will draw in religious and spiritual leaders, and practitioners in the arts. Our activities will be attended by representatives of the quality media, encouraging them to join in the task of understanding the technology and its implications for every aspect of human nature.

The project should lead to a widening consciousness of the significance of future AI for two distinct groups. In the first place, we would expect that our initiatives will lead AI researchers to initiate and maintain discussions with specialists in ethics, the arts, cultural studies, and philosophy of mind and religion, along with theologians and religious leaders. Secondly, we expect humanities groups to keep abreast of AI developments, and to contribute to debates and qualitative critiques.

[1] This was a key topic of discussion at the 2016 conference on Superintelligence and Humanity run by the SHDP Director at Jesus College, Cambridge. An executive summary is available at
www.jesus.cam.ac.uk/articles/machine-superintelligence-and-humanity-rustat-conference-report

Acknowledgements

The AI and the Future of Humanity Project runs for two years from August 2017. It is an initiative of the Science & Human Dimension Project, based at Jesus College, Cambridge since 1990. We thank the Master and Fellows of Jesus College, Cambridge and Templeton World Charity Foundation (TWCF) for their support of this project. We also thank all our speakers, chairs and helpers. This report was written by Rob Hart.

Conference Conveners

John Cornwell Director, Science & Human Dimension Project
Jonathan S. Cornwell Executive Director, Science & Human Dimension Project

Conference 1
Who's afraid of the Super-Machine? AI in Sci-Fi Literature and Film
Date 15-16 March 2018 **Location** Jesus College, Cambridge
Conference Overview
The term "singularity" was introduced by the science fiction writer Vernor Vinge in a 1983; it was picked up by Ray Kurzweil in his popular 2005 book The Singularity is Near. At many stages we find fiction in all its forms driving ideas in AI and vice versa. Crucially, we find the relationship between AI developments and our hopes, fears and ambitions, worked out imaginatively through a variety of media. Hence film and literary fictions have been a forum for the drama of ideas that circulate around AI and its future, not least its moral dimension. What can we learn about ourselves in relation to AI by exploring these narratives? There are also powerful religious themes in the history of SF machine intelligence, such as achievement of immortality, notions of Omega point futures, transhumanism, and the prospect of androids outstripping humans in virtue. Film makers, SF authors, researchers in literature, film studies, philosophy and the humanities addressed these questions with AI experts.

Conference 2
The Singularity Summit: Imagination, Memory, Consciousness, Agency, Values
Date 26-27 September 2018 **Location** Jesus College, Cambridge
Conference Overview
Numerous research projects around the world are attempting to simulate human "intelligence" based in part on neurophysiological theories of memory and imagination. Although considerable work has been done in this area since the early 1990s, AI is currently experiencing a quantum shift, one that requires an in-depth review of the primary human faculties as well as the moral dimension of human existence. While these research and development programs would benefit greatly from dialogue with philosophy of mind, aesthetics, literary and cultural studies, and philosophical theology, there is a lack of dialogue between scholars in these areas and the AI communities. This conference will provide a much-needed opportunity for interdisciplinary discussion between these groups who do not often find themselves around the same table. An important segment of the conference would involve standing back to review the future of AI in the historical context of how the digital age has already affected society and individuals.

Conference 3
Will advances in machine intelligence serve to enhance or diminish our moral and spiritual selves? Will these advances serve to create better or worse societies?
Date 16-17 May 2019 **Location** Jesus College, Cambridge
Conference Overview
In this conference we ask what impact future AI is likely to have on notions of the soul, religious faith, religious practice, and the virtues. Does AI pose a threat, or encouragement, to religious belief and practice, and will it create better or worse societies? In turn, we ask how religion might guide and inform attitudes towards, and relationships with, future intelligent machines. Finally, can religious perspectives influence and shape the course of AI research and development?

SHDP Science & Human Dimension Project

Science and Human Dimension Project is a public understanding of science, technology, engineering, maths and medicine (STEMM) program, based at Jesus College, Cambridge and founded in 1990. Through conferences and publishing, SHDP brings together the scientific research community with experts from industry, government and the media to deepen and broaden the appreciation of new ideas and discoveries and to ask searching questions about their impact on humanity.

SHDP addresses important ethical questions, such as the controversy over human embryonic stem cell research, superintelligent machines and artificial intelligence (AI). At times we are intent on tackling subjects illustrative of knowledge purely for its own sake.

SHDP helps university research, companies and think thanks bring their ideas to a wide audience of experts, the media and general public through carefully organised conferences with hand-picked participants. In 2017 we collaborated with DeepMind to deliver a conference on their AI research into memory and imagination.

AI & the Future of Humanity Project - from August 2017 we are running a two-year project of three conferences and related outreach on AI and the Future of Humanity, funded by Templeton World Charity Foundation (TWCF).

SHDP and outreach - we have a strong track record in achieving outreach from the university and the laboratory to the media and a wider non-specialist public through journalism, film and books. SHDP directors have produced conferences and reports on a wide range of vital issues including: Consciousness and Human Identity, AI, Cyber Security, Food Security, Inequality, Infrastructure, the Financial Crisis, Big Data, Blockchain and Bitcoin, the Future of Research-based Universities, the UK North-South Divide, Ageing, and the Future of Work. Conference proceedings and books have been published by OUP, Bloomsbury, Penguin and Profile.

Science and Human Dimension Project - the project is run by SHDP founder John Cornwell (Director) and Jonathan S. Cornwell (Executive Director). Advisors on the *AI & the Future of Humanity* project include Rev'd Dr Andrew Davison (Starbridge Lecturer, University of Cambridge), Rev'd Dr Tim Jenkins (Anthropologist, Fellow, Jesus College, Cambridge), Rev'd Dr Paul Dominiak (Theologian, Fellow, Jesus College, Cambridge), and Dr Tudor Jenkins (technologist and Artificial Intelligence researcher).

SHDP Science & Human Dimension Project
Jesus College, Cambridge

Tel: +44 (0)7768 220188 E: j.cornwell@jesus.cam.ac.uk www.Science-Human.org @ScienceHumanDP

Singularity Summit

Science & Human Dimension Project
Jesus College, Cambridge
26-27 September 2018

Speaker Bios and Abstracts

Dr Ron Chrisley

Dr Ron Chrisley is director of the Centre for Cognitive Science (COGS) at the University of Sussex, where he is also on the faculty of the Sackler Centre for Consciousness Science, and Reader in Philosophy in the School of Engineering and Informatics. He was awarded a Bachelor of Science from Stanford University and a DPhil in Philosophy from the University of Oxford. Before arriving at Sussex he was an AI research assistant at Stanford, NASA, RIACS, and Xerox PARC, and investigated neural networks for speech recognition as a Fulbright Scholar at the Helsinki University of Technology and at ATR Laboratories in Japan. From 2001-2003 he was Leverhulme Research Fellow in Artificial Intelligence at the School of Computer Science at the University of Birmingham. He is one of the co-directors of the EUCognition research network, and is an Associate Editor of Cognitive Systems Research and Frontiers in Psychology (Consciousness Research). He is also the editor of the four-volume collection Artificial Intelligence: Critical Concepts.

Abstract: In defence of the possibility of artificial consciousness: Computation, diagonalisation, and the qualia delusion

One argument against the possibility of artificial consciousness is the diagonalisation argument (put roughly: there are questions about computation that conscious humans can answer that no computer can; therefore no computer can fully behave like a conscious human; but fully behaving like a conscious being is a requirement for being a conscious being, so no computer can be conscious). Another is the argument from qualia (put roughly: all consciousness has qualitative aspects called qualia, qualia are not physical, computers are purely physical, so a conscious computer is impossible). I argue that neither of these arguments give us good reason to doubt the possibility of artificial consciousness. Concerning the first argument, I show that even if the first premise is granted, the rest does not follow: for one thing, X need not compute the same functions as Y in order for X to perfectly simulate Y. Concerning the second argument, I cast doubt on the assumption that qualia are necessarily non-physical by making intelligible a scenario in which the term "qualia" refers to physical phenomena, despite the fact that the properties typically ascribed to qualia are incompatible with them being physical. In such a scenario, a robot could have qualia by virtue of being in the same kinds of virtual machine states that, in us, explain/give rise to our (largely mistaken, though successfully referring) qualia beliefs and qualia talk.

Dr Robert Clowes

Robert Clowes is a Researcher and Invited Assistant Professor at the New University of Lisbon, Portugal. He also directs the Lisbon Mind & Reasoning Group at the New University of Lisbon, which specializes in research devoted to mind, cognition, and human reasoning. Robert's research interests span a range of topics in philosophy and cognitive science, including the philosophy of technology, memory, agency, and the implications of embodiment and cognitive extension for our understanding of the mind and conscious experience. He is particularly interested in the philosophical and cognitive scientific significance of new technologies, especially those that are built on top of global digital networks, such as the Internet and Web. His work has appeared in a variety of journals, including Review of Philosophy and Psychology, AI & Society, Phenomenology and the Cognitive Sciences, Philosophy and Technology, and the Journal of Consciousness Studies.

Abstract: Agency, Memory and the Cognitive Ecology of AI

Abstract: Human beings are sometimes said to be strong agents. That is, we are reflective, planful beings that (at least some of the time) regulate and govern ourselves through our plans and policies (Bratman, 2000). I have argued that strong agency is more involved with our use of artefacts than we often notice (Clowes, 2018 Online First). Yet, it is sometimes argued that this sort of agency is under threat from the 'smart' internet technology that constitutes an increasingly omnipresent part of our environment (Loh & Kanai, 2015). If the threat is real, we have to ask what is it about this technology that makes it such a challenge to agentive capabilities? In this talk, I will suggest that the smart technology now embedded in the internet need not undermine human agency. But this depends upon the design, implementation and ultimately the values that inform the design and implementation of this technology.

Bratman, M. (2000). Reflection, planning, and temporally extended agency. *The Philosophical Review, 109*(1), 35-61.
Clowes, R. W. (2018 Online First). Immaterial Engagement: Human Agency within the Cognitive Ecology of the Internet. *Phenomenology and Cognitive Science*.
Loh, K. K., & Kanai, R. (2015). How has the Internet reshaped human cognition? *The Neuroscientist*, 1073858415595005.

Professor Simon Colton

Simon Colton is a Professor of Computational Creativity at both Queen Mary University of London and Monash University, and was until recently an EPSRC leadership fellow and held an ERA Chair. He is an Artificial Intelligence researcher specialising in the study of Computational Creativity, where the aim is to build software which can take on creative responsibilities in arts and science projects. He is best known for writing software such as: The Painting Fool, where the aim is for the program to be taken seriously as an artist in its own right one day; HR which has made a number of discoveries in pure mathematics; The WhatIf Machine which performs fictional ideation for cultural purposes; and the Wevva app for semi-automated videogame design. These practical applications, along with substantial public engagement, have enabled a holistic overview of the notion of automated creativity, and Prof. Colton has contributed to theoretical frameworks for evaluation and philosophical discourse covering topics such as creative autonomy, framing and authenticity. He is currently working on automating creative aspects of computer programming.

Abstract: Should an AI cause insult and harm?

Coming from the perspective of Computational Creativity, I am interested in how generative AI software, which creates artefacts of value such as theorems, pictures or poems, is held back by a lack of authenticity, as it doesn't have a presence or life experiences in human terms in the human world. This leads on to questions of whether software can disrupt human practice, insult people and cause harm if it doesn't have an authentic voice. We hypothesise that for the software to truly cause insult (as opposed to the programmer/user or the words themselves that it generates), it needs both authenticity and creativity. Investigating how software can cause harm is, I believe, imperative, given that everyone from politicians and religious leaders to tech entrepreneurs and science fiction writers are claiming that AI will cause great harm to humanity, largely based on speculation rather than reasoned debate or empirical evidence.

John Cornwell - Director, Science & Human Dimension Project

After 12 years on the editorial staff of *The Observer*, he was in 1990 elected Senior Research Fellow and Director of the *Science and Human Dimension Project* at Jesus College, Cambridge. In that role he has brought together many scientists, philosophers, ethicists, authors and journalists to debate a range of topics in the field of public understanding of science. His edited books include *Nature's Imagination*, *Consciousness and Human Identity*, and *Explanations* (Oxford University Press); *Power to Harm,* and *Hitler's Scientists* (Viking Penguin); *The Philosophers and God* (Bloomsbury Continuum), and *Darwin's Angel* (Profile). He is a Fellow of the Royal Society of Literature and was awarded an Honorary Doctorate of Letters (University of Leicester) in 2011. He was shortlisted Specialist Journalist of the Year (science writing in *Sunday Times Magazine*), British Press Awards, 2006. He won the Science and Medical Network Book of the Year Award for *Hitler's Scientists*, 2005; and received the Independent Television Authority-Tablet Award for contributions to religious journalism (1994).

His journalism has been published in *Financial Times, Sunday Times Magazine, The Observer, New Statesman, New Scientist, Nature, Prospect, Times Literary Supplement, The Tablet, Brain, The Guardian, The Times.* Broadcast contributions to many BBC programmes, including "Hard Talk", "Choice", "Start the Week," "The Moral Maze", "Today" (debate with Richard Dawkins); "Beyond Belief", "Thought for the Day", "Sunday", and the BBC's World Service.

Abstract: Imagination and Memory

Writing in 2017 Demis Hassabis, CEO and co-founder of DeepMind, declared: "Imagination is one of the keys to general intelligence, and also a powerful example of neuroscience-inspired ideas crossing over into Artificial Intelligence." In our second conference, in a series of five, the Science & Human Dimension Project, attempted to explore in what sense the AlphGo system has a form of imagination; how that function might be compared and contrasted with what we understand by human imagination in its varied manifestations. That meeting brought together constituencies of AI specialists and people in the Humanities, including philosophy, psychology, and anthropology. To maintain a sense of continuity between our conferences, I shall give an overview of the thinking and discussions prompted by that meeting, with a suggestion for further exploration.

Dr Andrew Davison

Andrew Davison is the Starbridge Lecturer in Theology and Natural Sciences at the University of Cambridge, and the fellow in theology at Corpus Christi College. He studied chemistry at the University of Oxford, followed by a DPhil biochemistry, before turning to the study of theology in preparation for ordination in the Church of England, with a further PhD following later, in mediaeval philosophical theology. He served a curacy in South East London and is currently the Canon Philosopher of St Albans Cathedral. Before his current position, he taught Christian doctrine at St Stephen's House, Oxford and later at Westcott House, Cambridge. He works on a range of topics at the interface of theology, philosophy and natural science. His recent work has been on the theological significance of the prospect of life elsewhere in the universe, mutualism in biology, cosmology and creation ex nihilo, evolution and traditions of divine exemplarity, and the role of the imagination in responding to the threat of climate change. Next year will see the publication of book with Cambridge University Press on the Platonic theme of participation as a structuring principle in Christian theology and metaphysics.

Abstract: The Promise of Analogy: Talking about Machine Learning Across Interdisciplinary Boundaries

Our meetings at Jesus College offer a remarkable opportunity for conversation between scientists working at the forefront of artificial intelligence and scholars in the arts and humanities. They also show that these conversations can sometimes be difficult, with tensions over whether we can recognise the use of existing language in new situations. The scholastic philosophy of the high Middle Ages has a good deal to to offer here, not least in its extensive discussion of the category of analogy, where a word is used neither in precisely the same fashion across different settings (which would be 'univocity'), nor in a completely different way, such that no real comparison is implied (which would be 'equivocity'). To talk univocally of creativity or ingenuity, or even of memory, in a machine learning setting—as if these were precisely the same as in human beings—may be to claim too much; to assume that we mean the terms equivocally, on the other hand, might be to dismiss the remarkable properties that are observed in these systems. Analogy treads the middle way, and from these discussions of language it invites us to address some more metaphysical questions. In particular, we can ask how it is that the comparable properties we observe in our analogous domains—the human being and the AI machine—come to possess these similarities.

Dr Daniel De Haan

Daniel De Haan is a Research Fellow at the Ian Ramsey Centre for Science and Religion at the University of Oxford. Before coming to Oxford he was a postdoctoral fellow working on the neuroscience strand of Templeton World Charity Foundation's Theology, Philosophy of Religion, and the Sciences project at the University of Cambridge. He has a doctorate in philosophy from the Catholic University of Leuven and the University of St Thomas in Texas. His research focuses on philosophical anthropology and the sciences, philosophy of religion, and medieval philosophy.

Abstract: Someone but not a Self

Must one think of oneself as being a 'self'? With an eye to historical, ethnographical, and other considerations, I shall argue we need to interrogate and disengage certain theoretical impositions on commonsense or manifest images of human lives which distort both our own self-knowledge and theoretical enquiries in cognitive neuroscience, philosophy, theology, and many other disciplines. I focus on and illustrate why the "self" is among these distorting theoretical impositions. I argue that a human is not a 'self,' but a 'someone'. A 'someone' is a developing and dependent rational animal, one who's biopsychosocial personal identity develops in a variety of ways throughout their unified life as a person or a someone. I conclude with the curious fact that what is peculiar to someone's like us is both our ability to define ourselves and our ability to contort ourselves to fit these false and true identities.

Dr Marta Halina

Marta Halina is University Lecturer in Philosophy of Cognitive Science at the University of Cambridge, Programme Director at the Leverhulme Centre for the Future of Intelligence, and Fellow at Selwyn College. She holds degrees in biology and philosophy and received her PhD at the University of California San Diego in 2013 before becoming a McDonnell Postdoctoral Fellow in the Philosophy-Neuroscience-Psychology Program at Washington University in St. Louis in 2013-1014. She joined the History and Philosophy of Science Department at the University of Cambridge in 2014. Her research includes work on nonhuman animal cognition and communication, mechanistic explanation, and the use of comparative methods in cognitive science. Her recent papers include "The Goal of Ape Pointing" (2018, PLOS ONE) and "Not Null Enough: Pseudo-Null Hypotheses in Community Ecology and Comparative Psychology (forthcoming in Biology & Philosophy).

Abstract: Machine Intelligence and the Capacity for Insight

In March 2016, Google DeepMind's computer program AlphaGo surprised the world by defeating the world-

champion Go player, Lee Sedol. Go is a strategic game with a vast search space (including many more legal positions than atoms in the observable universe), which humans have been playing and studying for over 3000 years. Watching the tournament, the Go community was struck by AlphaGo's moves - they were surprising, original, "beautiful", and extremely effective. The moves were described as "creative" by the Go community and in follow up talks on the subject, Demis Hassabis - leading AI developer and CEO of Google DeepMind - defended them as such. Should we understand AlphaGo as exhibiting human-level insight? Answering this question requires having an account of what constitutes insightful thought in humans and developing tests for measuring this ability in nonhuman systems. In this talk, I draw on research in cognitive psychology to evaluate contemporary progress in AI, specifically whether new programs such as AlphaGo are best understood as exhibiting insight.

Dr Erik Myin

Erik Myin is Professor of Philosophy at the University of Antwerp and director of the Centre for Philosophical Psychology. He has published papers on topics relating to mind, experience and cognition in philosophical, interdisciplinary and scientific journals. Two books, *Radicalizing Enactivism: Basic Minds without Content*, and *Evolving Enactivism: Basic Minds Meet Content*, written by Dan Hutto and Erik Myin, were published by MIT Press in 2013 and 2017. In these papers and books, the view is defended that experience and cognition fundamentally are, and should be understood in terms of historically established interactions of organisms with environments. Contentful representation and computation, instead of forming basic explanatory tools, should be explained themselves as capacities that have gradually evolved in a context of sociocultural interaction.

Abstract: RECtifying Intelligence: Radical Enactive/Embodied approach to Cognition and Intelligence

In two recent books Daniel Hutto and myself have defended REC, or the Radical Enactive/Embodied approach to Cognition and Intelligence. According to REC, basic cognition is conceived of as dynamically unfolding embodied interaction with worldly offerings. Cognition, including perception, is intentional, in that it targets specific objects and aspects of the environment, but it does so without involving contentful representation or computation. REC does not deny that content involving or computational cognition exists—but only when cognition is scaffolded by normative shared practices such as the use of a language.

In my talk, I will present REC's arguments for challenging both the representational and the computational pillar of traditional views of cognition and intelligence. I will show how going the REC way undermines very influential arguments for the multiple realizability of intelligence.

Antonio Ramos Diaz

Originally an engineering student, Antonio Ramos Díaz obtained a BA in History and Philosophy from the University of Puerto Rico, Rio Piedras Campus. He went on to pursue graduate studies first at the University of St. Andrews (M.Litt.) and then at the University of Leuven (M.Phil.). He is currently a PhD candidate at the University of Leuven and is expected to defend his doctoral dissertation in November 2018. The title of his dissertation is 'The Kripke-Ross Argument Against Naturalizing Formal Understanding'.

Abstract: Whatever mathematical thinking or understanding is, can it be the same as what a physical computing mechanism does when it is said to compute or 'understand' a mathematical function?

The paper examines an argument against the claim that any physical computing mechanism *can* realize and duplicate formal (i.e. mathematical and logical) understanding. The most famous case against computational accounts of formal understanding is referred to as the Lucas-Penrose Gödelian argument. There is, however, an altogether different case that can be made against computationalism which was first developed in a series of unpublished lectures by Saul Kripke. Kripke's argument levels a more fundamental attack on computational accounts of mathematical and logical understanding and stands or falls independently of any Gödelian argument. This talk presents a version of Kripke's unpublished argument, briefly compares it to the Lucas-Penrose Gödelian argument and suggests that Kripke's argument offers the strongest case for the claim that formal understanding and activity cannot be the same thing as that which a physical computing mechanism does (or can do) when it is said to realize formal operations and compute mathematical functions.

Professor Kathleen Richardson

Kathleen Richardson is Professor of Ethics of Culture of Robots and AI at the Centre for Computing and Social Responsibility (CCSR) at De Montfort University. Kathleen completed her PhD at the Department of Social Anthropology, University of Cambridge and she has carried out research on different kinds of robots including social robots, robots for children with autism and sex robots. She is author of An Anthropology of Robots and AI: Annihilation Anxiety and Machines (2015), Challenging Sociality: An Anthropology of Robots, Autism and Attachment (2018). and Sex Robots: The End of Love (forthcoming 2019). She is also founder of the Campaign Against Sex Robots.

Abstract: Corporate individualism, robots and AI: A call to Martin Buber's I and Thou

Philosopher Martin Buber said the word I was a combined word – I was combined with *either* it (I-It) or Thou (I-Thou). The philosopher René Descartes said 'I think therefore I am' (Cogito, ergo sum). Buber's philosophy is an interpersonal paradigm. Descartes philosophy is a radical split between the self and other. Where does the project of making robots and AI sit between these two outlooks? And why does the difference matter to what it means to be human? This talk will explore these themes.

Dr Anders Sandberg

Anders Sandberg's research at the Future of Humanity Institute centres on management of low-probability high-impact risks, estimating the capabilities of future technologies, and very long-range futures. Anders is a Senior Research Fellow on the ERC UnPrEDICT Programme and the FHI-Amlin Collaboration. Topics of particular interest include global catastrophic risk, cognitive biases, cognitive enhancement, collective intelligence, neuroethics, and public policy. He is research associate to the the Oxford Uehiro Centre for Practical Ethics, and the Oxford Centre for Neuroethics. He is on the advisory boards of a number of organisations and often debates science and ethics in international media. Anders has a background in computer science, neuroscience and medical engineering. He obtained his Ph.D. in computational neuroscience from Stockholm University, Sweden, for work on neural network modelling of human memory.

Abstract: Definitions and Models of the Technological Singularity

Abstract: The concept of technological singularity is widely used in imagining the future, but has multiple components and meanings, some very dissimilar. This talk will look at the history and definitions of technological singularity, as well as how people have attempted to model or predict it.

Dr Markus Schlosser

Markus Schlosser is a lecturer at the School of Philosophy, University College Dublin. His main research area is the philosophy of action, and he has worked extensively on issues in the metaphysics of agency, moral psychology, mental causation, the problem of free will, and the neuroscience of free will. His work is published in journals such as Philosophical Studies, Analysis, Synthese, the Journal of Ethics, and various collections and editions. In his recent work he has been working on embodied cognition and representationalism and he has become interested in Eastern perspectives on agency and the self, in particular from Buddhism and Advaita Vedānta.

Abstract: Agency without agents

The causal theory of action has been the standard view in the contemporary philosophy of mind and action. It says, roughly, that agency is to be explained in terms of causation by mental states (desires, beliefs, and intentions). A frequently voiced objection is that this view 'leaves out the agent': the standard theory reduces agency to a nexus of causal pushes and pulls in which no one (no agent) is doing anything. The debate about this presumes, firstly, that there is genuine agency and, secondly, that agency requires the participation of agents. Inspired by Eastern philosophy, I will turn around the dialectic and question the second presupposition. On Eastern views, the sense of being an individual self (or agent) is an illusion. There is agency, but, metaphysically speaking, there is no one doing anything. I will attempt to defend this position and in the final part I will briefly consider the implications for the prospects of genuine AI. I will suggest that we need to consider, in addition, the role of consciousness in agency. If the role of consciousness can be reduced to the role of mental states, then the prospects for genuine AI seem very good. But if consciousness is irreducible, genuine AI may be impossible.

Dr Simone Schnall

Simone Schnall is a Reader in Experimental Social Psychology at the University of Cambridge, and Director of the *Cambridge Body, Mind and Behaviour Laboratory*. She previously held appointments at Harvard University, the

University of Southern California and the University of Virginia. Her research explores why people often think and behave in seemingly surprising ways, and how to capitalize on insights from behavioural science to encourage adaptive choices in everyday life. Current topics include judgments and decisions in moral and legal contexts, perceptions of the spatial environment, and risky behaviours in finance. Her work has been supported by numerous grants from public and private funding sources, and is routinely covered in the popular media such as the New York Times, The Economist, Newsweek, New Scientist and Psychology Today. Dr. Schnall is a Fellow and Director of Studies for Psychology at Jesus College, Einstein Fellow at the Berlin School of Mind and Brain at Humboldt University, and a Fellow of the Association for Psychological Science.

Abstract: Embodied Perception and Action – A Challenge for AI?

Traditionally, perception has been considered to be an objective reflection of physical properties of the environment, such as an object's texture or geometrical shape. Computational approaches to cognition furthermore assume that while sensation and perception are modality-specific, higher-order thought processes are amodal and abstract. In contrast, embodied and enacted approaches propose that perception is a reflection of subjectively experienced possibilities for action (i.e., affordances). I will review experimental evidence to suggest that visual perception of the spatial environment takes into account the current bodily state of the perceiver, and relevant action goals. For example, a person's blood glucose level or their social resources influence how they see a hill. For disembodied intelligence a lack of, first, sensorimotor input, and second, affordances for action, may pose a significant challenge that will be critical to address.

Fr Ezra Sullivan OP

Ezra Sullivan is a Dominican friar and professor of Moral Theology and Bioethics at the Pontifical University of St Thomas Aquinas (Angelicum) in Rome, where he conducted doctoral research with Wojciech Giertych, the Theologian of the Papal Household. His articles and reviews have been published in such journals as Linacre Quarterly, Nova et Vetera, and Logos Journal. He is helping develop the Humanity 2.0 Project, and is frequently invited as a guest on Catholic radio and television. His forthcoming book is entitled Habits and Holiness: Thomistic Virtue Theory in Light of Modern Science (2019).

Abstract: Robots without Families: On Identity and Organic Continuity

Summary: Whereas initial stages of Artificial Intelligence research focused on how to make robots emulate human behaviours so closely as to be indistinguishable from human counterparts, fashioning them to be able to pass the Turing Test, later stages have striven to given robots a greater "interiority" to approximate that of humans. Thus, AI research into machine learning has attempted to give robots powers to advance themselves by learning through time. Similarly, products such as artificial skin seem to confer abilities to have sensations. One element that seemingly cannot be conferred is an identity based on organic continuity with others. In other words, robots cannot have families.

A vital element of human identity is one's location within a family. On the immediate level, this means sharing a genetic code with a mother and a father. Further within the family tree are siblings, cousins, uncles, aunts, grandparents, and so on, who share family traits that condition one's health, looks, behaviours, likes, and dislikes. On a wider scale, inheritance studies suggest that all living humans are genetically linked to a few common ancestors. Even further back, evolutionary studies indicate that all life on Earth can be traced back to a common material starting point. Consequently, human identity is bound up with organic continuity with past humans and all other living things. We might call this the "Family of the Living." As products of the choices of a rather small number of modern programmers, manufacturers, capitalists, and so on, robot physical identity is necessarily excluded from the evolutionary-biological family. The paper would explore some consequence of this fact for robot identity as a whole and their self-consciousness, if such a state is ever achieved.

Professor Steve Torrance

Steve is a visiting senior research fellow in the Centre for Cognitive Science (COGS) at the University of Sussex, where he has also been, at different periods, an external examiner, and a visiting lecturer in AI and cognitive science, and where he completed an undergraduate degree in philosophy 50 years ago. He has a DPhil from Oxford on the formal structure of ethical judgment, and he has been working on the relation between AI, philosophy of mind and ethics since the early 1980s. He is Emeritus Professor of Cognitive Science at Middlesex University, where he worked in the philosophy, computing and psychology departments. He has organised many conferences and workshops, and has edited collections and journal issues, on topics including philosophy of AI, machine ethics, machine consciousness, social robotics, and enactivist theory. He writes on the moral status of AI agents; limits to computationalism as a

theory of mind and consciousness; enactivist conceptions of mind and interaction; and the implications of technology growth for human civilisation. Working with a fellow jazz pianist and cognitive scientist, he also presents lecture-performances and writes on jazz and improvisation. He is an ethics reviewer for the European Commission's Horizon 2020 programme.

Abstract: Explanatory and Performance Gaps in Artificial Super-Intelligence

From a philosophical perspective AI is often considered to go hand-in-hand with a pan-computational view of mind, that explains everything mental in humans and other creatures in brain-centred, information-processing terms. Whether such a theory is adequate is now important for practical as well as theoretical reasons: many are now claiming that the advent of artificial super-intelligence (ASI) is not far in the future. And it is natural for supporters of the practical ASI project also to support a variant of a brain-based informational theory of mind.

Opponents of such a theory have pointed to possible shortcomings or gaps in this explanatory model. Can a merely informational or cognitive AI model of mind adequately explain phenomenally conscious states, emotions, intrinsic teleology, embodiment, sensorimotor interaction, not to mention the precarious and continually self-constituting (autopoietic) nature of the existence of a biological creature with mind? Any explanatory gaps in the informational account of mind will probably translate, in practical engineering terms, into capability gaps in fully-developed ASIs. To take one banal example that permeates our lives – food: many aspects of our alimentary physiology, and how the latter subserves humans' cradle-to-grave experience of eating and drinking, may be difficult to implement in a non-biological ASI, other than through baroque work-arounds. Given the many physical and organisational disparities between humans and digital ASIs, there could be many other such constitutive differences which could result in gaps in the efficacy of ASI performance. Given the apparent potentiality for ASI to dominate and radically reshape human civilization, the possibility of significant performance gaps – particularly if they are not obvious or easy to compensate for – should be widely debated before any programme for accelerating the development of super-AI is licensed to go ahead.

Abstract: Ethics for E-persons

In February 2017 a European Parliament Resolution on Civil Law Rules of Robotics, proposed by Luxembourg MEP Mady Delvaux, recommended to the EC that there should be a specific legal status for advanced autonomous robots, and that these should be regarded as "electronic persons" which may carry moral or legal responsibility for the consequences of their autonomous decisions. (This may include, for example, injuries and deaths caused by driverless vehicles.) Such a recommendation has been opposed in an Open Letter by a group of AI researchers across the EU (http://www.robotics.openletter.eu). The Letter argues that such an "electronic personality" status for robots, if treated as genuinely autonomous, would open up the possibility that certain robots should enjoy rights conferred to humans under the EU Charter of Fundamental Rights – such as the rights to dignity, integrity of the person, property, and so on.

It could be argued, against the Delvaux resolution, that moral responsibility in an agent presupposes a kind of phenomenal consciousness in that agent that electronic persons could never achieve because of their lack of organic physiology. Advanced, functionally autonomous robots, while they may have many person-like properties, might not pass any test for phenomenal consciousness likely to command wide scientific assent. However, this may not be sufficient to invalidate every such policy in the area of E-personality and moral status (both rights and responsibilities). I will discuss cases where a non-conscious E-person may be given rights to own property. These rights may in turn impose on humans moral obligations to accord such owners the same rights against arbitrary expropriation of their possessions as human property owners enjoy. Such cases may have far-reaching implications that deserve careful consideration.

Dr Karina Vold

Karina Vold is currently a postdoctoral research associate at the Leverhulme Centre for the Future of Intelligence, a research fellow at the Faculty of Philosophy at the University of Cambridge, and a Canada-UK Fellow for Innovation and Entrepreneurship. She received her PhD in Philosophy from McGill University and her BA from the University of Toronto. Her current work focuses on theories of cognitive extension, intelligence augmentation, consciousness, and AI ethics.

Abstract: Brain-computer interfaces: new varieties of cognitive imbalance

Driven by investments from military organizations and large tech companies, and coupled with recent advancements in deep learning, a race is emerging to develop cognitive enhancements through brain-computer interfaces (BCIs).

BCIs allow for a direct communication pathway between the brain and a computing device and, hence, their possibility raises many socio-ethical questions. In this talk I will attempt a preliminary typology of BCIs and will consider the range of cognitive capacities they are aimed at enhancing. I argue that BCIs have the risk of bringing about new kinds of cognitive imbalances in humans and, hence, carry the potential of distorting the human mind as much as enhancing it.

Science & Human Dimension Project - SHDP

Jonathan Cornwell - Executive Director SHDP
Jonathan has a background in academic, educational and digital publishing in Europe, the Middle East and China. From 2010-17 he was co-director of the Rustat Conferences at Jesus College, Cambridge, producing conferences that brought together academic experts with leaders from government, industry and the media, and he is currently Executive Director of the Science & Human Dimension Project (SHDP) and a Senior Research Associate at Jesus College, Cambridge. He has produced conferences and edited reports on a wide range of topics including Artificial Intelligence, cybersecurity, blockchain, bitcoin, energy security, food security, inequality, north-south divide, the future of work, and ageing. He also works with curators, artists, and galleries to produce exhibitions, including *Houghton Revisited: Masterpieces from the Hermitage*; *James Turrell: Lightscape*; *Beyond Beauty: Transforming the Body in Ancient Egypt*; and Chris Levine, *Inner DeepSpace*. He studied at UCL, Trinity Hall, Cambridge and Imperial College London.

Rob Hart - Conference Rapporteur, SHDP
Robert David Hart is a London-based journalist and writer with interests in science, technology and health, with particular interests in synthetic biology, robotics and artificial intelligence. His work has featured in, amongst other venues, Quartz, Slate, Wired and the Guardian. He has a bachelor's degree in natural sciences and a master's degree in the history and philosophy of science, both from Downing College, Cambridge.
To contact the rapporteur please write to j.cornwell@jesus.cam.ac.uk

Dr Tudor Jenkins - Adviser, SHDP
Tudor Jenkins' background is in Artificial Intelligence and Artificial Life, where he used a situated adaptive behavioural approach to understand conditions under which grammars can evolve in social agents controlled by neural networks. Tudor is currently interested in how adaptive systems can be better used to understand and improve human performance in sport. He holds a PhD in AI from Sussex University and did research at the École Normale Supérieure, Paris. As director of Wide Eyed Vision, he consults on digital aspects of cultural heritage and exhibitions. He is an adviser to Science & Human Dimension Project.

Colin Ramsay Director, DragonLight Films – director and producer with special interest in AI. Colin filming interviews at the conference. https://dragonlightfilms.com To view filmed interviews from the conferences visit the Science & Human Dimension Project website.

Dr Elisabeth Schimpfössl - assisting SHDP
YJ Gahng Judge Business School, University of Cambridge - assisting SHDP
Sam Thorp - assisting SHDP
Sam Thurman - assisting SHDP

Science & Human Dimension Project
Jesus College
Cambridge
CB5-8BL
j.cornwell@jesus.cam.ac.uk
www.science-human.org

Singularity Summit **Participants List**

Prof Jean Bacon	Emeritus Professor in Distributed Systems, Fellow in Computer Science	Jesus College, Cambridge
Rhys Blakely	Science Correspondent	The Times
Prof Tim Bliss FRS	Neuroscientist	Francis Crick Institute
Prof Margaret Boden	Research Professor of Cognitive Science	University of Sussex
Andrew Brown	Journalist, author, editor	The Guardian
Jordan Burgess	Research Scientist	Amazon
Dr Fenella Cannell	Associate Professor, Department of Anthropology	LSE
Geoff Carr	Science Editor	The Economist
Dr Yaqub Chaudhary	Templeton Fellow in Science & Religion	Cambridge Muslim College
Dr Ron Chrisley	Reader, Cognitive Science	University of Sussex
Prof William Clocksin	Dean, School of Computer Science	University of Hertfordshire
Dr Robert Clowes	ArgLab, Nova Institute of Philosophy - IFILNOVA	Universidade Nova de Lisboa
Prof Simon Colton	Professor of Computational Creativity, Artificial Intelligence and Games, Game AI Research Group, Dept. of Electronic Engineering & Computer Science	Queen Mary University of London, and Monash University
Prof Alastair Compston	Professor Emeritus of Clinical Neurosciences, Department of Clinical Neurosciences; Emeritus Fellow, Jesus College	University of Cambridge
Jonathan S. Cornwell	Executive Director, Science & Human Dimension Project	Jesus College, Cambridge
John Cornwell	Director, Science & Human Dimension Project	Jesus College, Cambridge
Sally Davies	Editor, Science & Technology	Aeon
Dr Andrew Davison	Starbridge Lecturer in Theology and Natural Sciences; Fellow in Theology, Corpus Christi College	University of Cambridge
Jeremy Dawson	Director of Photography	DragonLight Films
Dr Daniel De Haan	Research Fellow, Ian Ramsey Centre for Science and Religion	University of Oxford
Dr James Dodd	Technologist and investor; Foundation Member, Rustat Conferences, Jesus College, Cambridge	
Justin Doherty		Hemington Consulting
Rev'd Dr Paul Dominiak	Dean of Chapel and Fellow; Dep Director, AI & Future of Humanty Project, Science & Human Dimension Project	Jesus College, Cambridge
Madeline Drake	Housing, Health and Social Policy Consultant	
Dr Haibo E		DeepMind Ethics & Society
Prof Paul Fletcher	Bernard Wolfe Professor of Health Neuroscience	University of Cambridge
Dr Iason Gabriel	Philosopher	DeepMind Ethics & Society
Yeonjean Gahng	Judge Business School; Assisting Science & Human Dimension Project	University of Cambridge
Esam Goodarzy	AI Initiative	Harvard Kennedy School
Joy Green	Senior Futures Specialist	Forum for the Future
Vadim Grigoryan	Creative and Art Director	
Dr Marta Halina	University Lecturer in the Philosophy and Psychology of Cognitive Science; Project Director, Leverhulme Centre for the Future of Intelligence; Fellow, Selwyn College	University of Cambridge

Rob Hart	Freelance journalist; rapporteur	Science & Human Dimension Project
Prof Julian Hughes	RICE Professor of Old Age Psychiatry; Philosopher	University of Bristol
Dr Julian Huppert	Director, Intellectual Forum	Jesus College, Cambridge
Rev'd Dr Tim Jenkins	Reader in Anthropology and Religion; Fellow, Jesus College, Cambridge	University of Cambridge
Dr Tudor Jenkins	Director, Wide Eyed Vision; AI Researcher; Advisory Board	Science & Human Dimension Project
Dr Duncan Kelly	Department of Politics and International Studies; Fellow, Jesus College, Cambridge	University of Cambridge
Dr James Lefanu	Doctor, journalist, historian of science and medicine	The Daily Telegraph
Dr Siân Lindley	Senior Researcher, Human Experience & Design (HXD)	Microsoft Research
Keith Mansfield	Publisher, novelist, TV writer and broadcaster	
Dr Christopher Markou	Researcher in Law and Technology; Faculty of Law; Jesus College	University of Cambridge
Cassius Matthias	Founder and Editor, Yes & No magazine; artist, film maker	
Dr Neil McBride	Reader in IT Management, School of Technology	De Montford University
Prof Michael McGhee	Hon Senior Fellow, Philosophy	University of Liverpool
Dr David Mellor	Senior Lecturer in Sociology	University of South Wales
Dr Ken Moody	Fellow in Computer Science	King's College, Cambridge
Oliver Morton	Briefings Editor; science writer	The Economist
Dr Ashley Moyse	McDonald Centre for Theology, Ethics, and Pubic Life	University of Oxford
Dr Erik Myin	Philosophy of Mind and Cognitive Science	University of Antwerp
Prof John Naughton	Senior Research Fellow, CRASSH; Fellow, Wolfson College; Emeritus Professor of the Public Understanding of Technology, OU; Technology Columnist, The Observer	University of Cambridge
Sumit Paul-Choudhury	Editor Emeritus	New Scientist
Sr. Dr. Alicia Pérez FCJ	Chaplain; AI researcher	Manchester Universities Catholic Chaplaincy
Blanca Pérez Ferrer	Cultural Mediator, Games Academy	Falmouth University
Marie-Therese Png	DeepMind Ethics & Society	DeepMind
Dr Antonio Ramos Diaz	Philosophy Researcher	University of Leuven
Colin Ramsay	Director, Producer	DragonLight Films
Prof Lord Martin Rees	Emeritus Professor of Cosmology and Astrophysics; former President, Royal Society; Astronomer Royal	University of Cambridge
Prof Kathleen Richardson	Professor of Ethics and Culture of Robots and AI	De Montford University
Prof Peter Robinson	Professor of Computer Technology	University of Cambridge
Prof Dale Russell	Visiting Professor, School of Design	Royal College of Art
Dr Anders Sandberg	Research Fellow, Future of Humanity Institute, & Oxford Martin Senior Fellow, Oxford Martin School	University of Oxford
Mary Savigar	Editor	Polity Press
Dr Elisabeth Schimpfössl	Lecturer in Sociology and Policy	Aston University
Dr Markus Schlosser	Lecturer in Philosophy	University College Dublin UCD
Dr Simone Schnall	Reader in Experimental Social Psycholody; Director, Cambridge Body, Mind and Behaviour Laboratory	University of Cambridge
Dr Abigail Sellen	Principal Researcher, Deputy Director	Microsoft Research

Prof Murray Shanahan	Professor of Cognitivie Robotics, Imperial College; Senior Research Scientist	DeepMind
Dr Beth Singler	Researcher "Human Identity in an age of Nearly-Human Machines" project, Faraday Institute; Research Fellow, Leverhulme Centre for the Future of Intelligence	University of Cambridge
Michael Stevenson	Senior Advisor, Education & Skills	OECD
Fr Ezra Sullivan OP	Theology Faculty, Pontifical University of St Thomas Aquinas	Angelicum, Rome
Sam Thorp	Assisting Science & Human Dimension Project	
Sam Thurman	Assisting Science & Human Dimension Project; Marketing Executive	Mojeek
Prof Steve Torrance	Centre for Research in Cognitive Science (COGS)	University of Sussex
Dr Cozmin Ududec	Head of Research	Invenia Labs Cambridge
Dr Liesbeth Venema	Chief Editor	Nature Machine Intelligence
Dr Karina Vold	Postdoctoral Researcher, Leverhulme Centre for the Future of Intelligence; Research Fellow, Faculty of Philosophy; Newnham College	University of Cambridge
Richard Watson	Futurist in Residence, Tech Foresight	Imperial College London
Dr Adrian Weller	Programme Director for Artifical Intelligence, The Turing Institute; Senior Research Fellow, Departent of Engineering	University of Cambridge
Dr Daan Wierstra	Senior Research Scientist	DeepMind
Dr Bonnie Zahl	External Advisor, Templeton World Charity Foundation	TWCF