Mr Antonio Rago

RECOMMENDER SYSTEMS

Professor Toni and her group are using argumentation for recommender systems such as those used by movie-streaming services. Antonio Rago, a PhD student in Professor Toni's group, says: "Recommendations are becoming more and more important these days. Users just don't have time to go through numerous reviews, so they allow systems to recommend items for them."

Mr Rago and fellow PhD student Oana Cocarascu are working with Professor Toni to develop a new kind of recommender system. In common with those currently in commercial use, their system uses a combination of collaborative filtering, which works by exploiting the fact that two users who both like one film have a better-than-average chance of agreeing in their assessment of another, and content-based filtering, which works by exploiting the fact that a user who likes one film is likely to enjoy other films which share the same features, for example its director, actors or genre.

What distinguishes this new system is that it is designed to represent the relevant data as a framework of arguments standing in dialectical relations of support or attack to one another. For example, one argument may support the conclusion that a user will like Catch Me If You Can because she likes the director, while another argument may attack this on the basis that she rated other films in the genre 'drama' poorly.

Representing the data this way has an important advantage: "Once the system has computed the score and given the user a recommendation, the user can actually respond to the system with, 'Why did you recommend this to me?'", Mr Rago explains. Since users understand the rationale, they have greater reason to trust the system and continue using it.

This is not the only advantage. "Once given this explanation, the user can basically say 'well this is wrong'", says Mr Rago. "If the argument representing the film's director had the biggest positive effect on this film's recommendation, the user may say: 'I don't care about directors – that's just not an aspect that's important for me'. Then every director in the graph will have a smaller effect on the films that they directed. Using this form of feedback, among many others that our system allows, we get iteratively better recommendations"

This is key to Professor Toni's conception of argumentation-based AI: "We need to work, if you like, in symphony with the AI", she says, "so it's not just the AI in command". This consideration is central to many of the applications in which Imperial's AI researchers are using argumentation to aid explainability.

CYBERSECURITY

Dr Erisa Karafili

Dr Erisa Karafili, in the Department of Computing, was recently awarded a prestigious Marie Curie Individual Fellowship to develop an argumentation-based AI tool for cybersecurity. The tool is intended to assist in the attribution of cyber-attacks, helping analysts understand the modus operandi of particular attackers and put prevention or mitigation measures in place.

To attribute a cyber-attack, one must understand not only the technical features of the attack, but also the geopolitical and social background. The system Dr Karafili is developing will take into account forensic evidence such as the attack's source IP addresses, and social evidence, for example the political motivations a country has for launching cyber-attacks.

In some cases, Dr Karafili says, the two kinds of evidence conflict: "The IP address might be telling you that the attack came from a country which does not have the motivation or capabilities to carry out an attack. This might suggest that the attack did not come from that country and the IP address is not reliable evidence."

Her tool is intended to make things easier for cyber-security analysts, who often have little time and a huge amount of evidence to work through. But because the tool will represent its conclusions and the considerations leading to them using an argumentation framework, the analyst's own background knowledge and acumen will not be wasted.

"We always want the analyst here", says Dr Karafili. "The analyst is the person who has the knowledge and it's their experience that directs them. The explainability provided by the argumentation framework means that they can check the system's preferences and decide which rules are stronger than others."

"The aim is for the tool to explain to the analyst why certain decisions were made, and permit the analyst to say 'no, I don't agree with how this decision was made, change it and give me another result'. In this event, the tool will also provide the analyst with insights into new paths for investigation."

MEDICAL DECISION-MAKING

One area that AI has potential to transform is healthcare, and this is reflected at Imperial by a variety of initiatives including the UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare led by Dr Aldo Faisal from the Departments of Computing and Bioengineering, the patient-centric UK Dementia Research Institute and the British Heart Foundation Centre of Research Excellence, which all aim to integrate state-of-the-art medical understanding and AI-led interventions.

Healthcare AI is also an area in which argumentation has promising applications, and these are being pursued by ROAD2H is an EPSRC-funded research project led by Imperial researchers, including Professor Lord Ara Darzi in the Faculty of Medicine and Professor Toni, in collaboration with researchers at King's College London, the University of Belgrade in Serbia and the China National Health Development Research Centre.

Dr Kristijonas Cyras

One researcher on the project, Dr Kristijonas Cyras in Imperial's Department of Computing, explains: "Lots of elderly people have multiple conditions such as hypertension, diabetes, and chronic obstructive pulmonary disease (COPD). For different conditions there are different treatment methods that can interact. While there are documented guidelines for treating individual diseases, for comorbid diseases there are, at best, only standard practices."

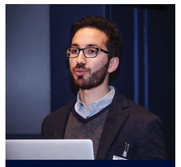
Dr Cyras continues: "For an individual human it's pretty hard to go over all those and make a decision, and doctors don't always have enough time." The AI system the project is developing can help. It takes multiple treatment recommendations, formalised in a computer interpretable way, models them, and, for a given patient, tells the user which recommendations apply and how they are likely to interact with each other.

It is, of course, crucial that the clinician can trust the system, and to this end the system offers reasons behind all the recommendations it makes, for instance reminding the clinician that a particular treatment is recommended by medical guidelines.

The human-style of reasoning employed by the AI has the further advantage of allowing humans to actively collaborate with the system. For example, the clinician and patient may decide it is more important to avoid exacerbating the patient's angina systems than it is to manage their COPD. If the clinician instructs the AI accordingly, the system will only recommend a COPD treatment unlikely to adversely affect the cardiovascular system. "This recommendation may change if you change your preferences, and the system will take a different decision on the fly and explain that as well", says Dr Cyras.

EXPLAINABLE ROBOTICS

It is not only disembodied machines that are driven by artificial intelligence. Engineers are increasingly harnessing the power of AI to build intelligent robotic systems that behave intelligently and even work collaboratively with humans, for example on factory assembly lines. For these to work effectively, humans must be able to understand and trust the robots. "The complaint that people tend to have with collaborative robots is that their behaviours are unpredictable," says Dr Ali Shafti, a research associate in the Department of Computing and team member in Dr Faisal's Brain & Behaviour Lab.



Dr Ali Shafti

The problem often arises, Dr Shafti explains, with robots that rely heavily on deep learning methods. The issue is not just that their algorithms are difficult to decode, but that the behavioural strategies the algorithms generate to help the robot achieve a goal may be different to the strategies familiar to humans.

Dr Shafti is addressing this problem by developing robots that use hierarchically-structured control algorithms that interact through reinforcement learning. At the top level of the hierarchy is an agent that plans its actions at a very high level of abstraction. Like a human, it does not know which joints need to be oscillated at what speeds, for example, it just knows that it needs to get from point A to point Z. It instructs lower-level agents to achieve sub-goals like getting from A to B, and those agents send instructions to agents below them, bottoming-out at the lowest level, generating behaviours that are very fine-grained and finely-sensitive to environmental detail.

"Because these low-level behaviours are governed by high-level action plans like those humans consciously use, we can co-ordinate our behaviours much more effectively with the robots," says Dr Shafti. As part of the recently concluded Horizon 2020 project eNHANCE, which aimed to improve robotic limbs that help paralysed people reach and grasp, he extracted human behavioural hierarchies in the form of action grammars, leading to a more intuitive human-robot interaction experience.

"If we can combine the hierarchical architecture used as part of that project with our hierarchical reinforcement learning methods," says Dr Shafti, "we will next be able to create a robotic system that not only helps you with reaching and grasping, but also learns the way you do things, learns your hierarchies, and over time allows you to learn the robot's hierarchies so you can work together in a better collaboration".

VARIETIES OF EXPLAINABILITY

Creating AI that is both powerful and explainable is a big challenge, encompassing a wide variety of sometimes incompatible theoretical perspectives. But Professor Toni is optimistic. She says: "Our researchers are pursuing some very varied and interesting interpretations of explainability. This means we're perfectly equipped to direct this important aspect of AI research, and shape the development of systems that, besides performing well, are understandable to experts and non-experts alike."

This feature is based partly on work presented at the 'explAIn' workshop at Imperial College London, 25 April 2018, organised by Professor Francesca Toni. Further information and references are available from the workshop page at **bit.ly/2yn3eFX**.

Dr David Silverman is a communications officer in Imperial's Enterprise Division. He was previously a philosopher of cognitive science.

www.imperial.ac.uk/artificial-intelligence

Imperial College London

EXPLAINABLE AI

Machines we can trust, learn from and collaborate with

BY DAVID SILVERMAN

Artificial intelligence (AI) systems can now perform a range of tasks that previously required the cognitive skills of humans. In some cases, they can even perform these tasks better or faster than we can. However, the principles that AI systems use to make intelligent decisions are often hidden from their end users.

Consider the systems that movie streaming services use to recommend films they think you personally will enjoy. The recommendations are rarely accompanied by explanations of why the systems have suggested those particular films. This is one reason we place less confidence in machine-generated recommendations than those from friends.

Some AI systems are so complex that even their designers do not fully understand the decision-making procedures they use. This can make it hard to be sure, for instance, that a self-driving car will drive safely even in situations it has not previously encountered, or that an algorithm a company uses to make decisions that affect consumers financially will do so without racial or sexual prejudice. Trust is particularly important when safety or fairness is at stake.

There is, therefore, a move toward the development of 'explainable' AI systems whose decision-making procedures are accessible to humans. Exploring the work of just some of the Imperial researchers working to develop explainable AI is enough to reveal the breadth of thinking on the topic. AI researchers from across the College's faculties are brought together under the AI@Imperial network, led by Dr Seth Flaxman, Professor Stephen Muggleton and Professor Francesca Toni. For some, the challenge is to create machines we can trust. For others, it is a question of creating machines we can learn from and even collaborate with in joint decision-making.

Professor Francesca Toni

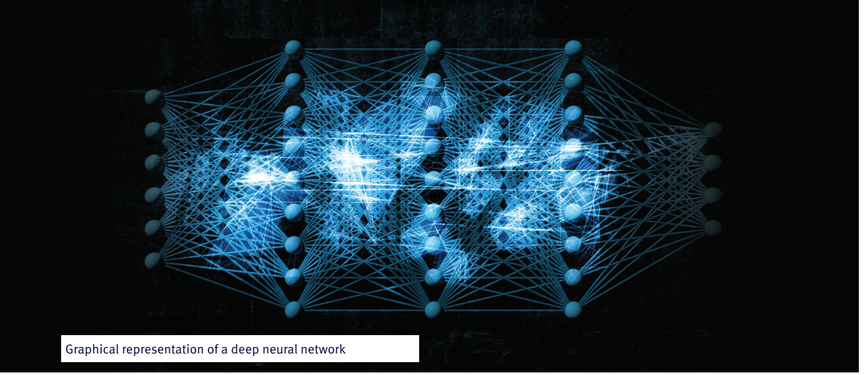
Professor Alessio Lomuscio

DEEP LEARNING

Much of the buzz currently surrounding AI is directed at machine learning – AI technology that learns through training to perform its tasks more effectively. While there are many approaches to machine learning available, the most popular today is arguably deep learning.

In deep learning, a machine employs a neural network: a set of interconnected nodes, in this case formed of tens or even hundreds of hierarchical layers, that partly resembles the network of neurons and synapses in a biological brain. The connections between the nodes can number in the millions, and have varying weights, meaning that a node's propensity to influence the behaviour of any other specified node can vary.

By learning appropriate weights through training on very large datasets, the system becomes sensitive to complex statistical relationships between variables, and in consequence gains an often impressive ability to classify noisy and ambiguous incoming data appropriately. This enables the system to interact intelligently with its users and environment.



Graphical representation of a deep neural network

Deep learning architectures are being widely used in applications such as speech and image recognition, because they are so effective at classifying data correctly. However, the complexity of the architecture means that humans cannot typically understand the principles the systems apply to make decisions. Professor Alessio Lomuscio, in Imperial's Department of Computing, puts the point starkly: "Deep learning systems are huge, deep and inaccessible – they operate like black boxes."

Imperial's AI researchers are responding to this problem in varying ways. While some are advocating the development of very different architectures, others are looking to better understand and verify the procedures already used by deep learning and other approaches that use neural networks.

VERIFYING AUTONOMOUS SYSTEMS

Professor Lomuscio is the Imperial lead for the UK Research and Innovation (UKRI) Centre for Doctoral Training in Safe and Trusted Artificial Intelligence, which brings together experts from Imperial and King's College London to train a new generation of researchers in AI and its ethical, legal and social implications.

He also leads the Verification of Autonomous Systems research group. With support from a grant from the Defense Advanced Research Projects Agency's Assured Autonomy programme, Professor Lomuscio and the group aim to provide guarantees that AI systems based on neural networks perform correctly even in safety-critical situations.

Autonomous cars are one important example. "Recently, a self-driving Uber car crashed and killed a pedestrian who was wheeling a bicycle. This happened in spite of the fact that its sensors picked her up several seconds before the accident", Professor Lomuscio says. "We need to ensure that behaviour like this doesn't happen."

To this end, his research group is beginning to develop mathematical techniques and toolkits for verifying the behaviour of neural networks. The aim is to analyse these neural systems before deployment and mathematically verify them. "This is not a matter of experimenting on them, but arriving at mathematical proofs that will give us complete confidence that safety critical systems will perform correctly in all scenarios", he explains. The techniques the team develops will not be restricted to particular AIs but will be applicable widely.

Dr Seth Flaxman

GDPR

Explainable AI has taken on particular significance with the advent of the General Data Protection Regulations (GDPR). When the rules became law in 2016, pending a two-year implementation period, Dr Seth Flaxman, now a lecturer of statistics in Imperial's Department of Mathematics, pointed out that the regulations potentially give individuals what he and a colleague called a 'right to explanation', a phrase that has since become widely used.

Dr Flaxman is now in high demand from businesses aiming to understand the implications of GDPR. "If we have a black box machine learning method, and someone comes along and demands to know why the method made a particular consequential decision – for example, why they were turned down for a home loan, or offered a particular deal on a holiday – that person can demand to know why", he explains. "We are not at the moment in a position where we can take black box machine learning methods like deep learning and answer that question in any meaningful way."

Dr Flaxman describes an existing algorithm whose decision-making procedures are theoretically easy to explain but raise all sorts of issues when we try to explain them in practice. On a blackboard, he draws two axes – one for age, and one for income – and dots at various co-ordinates. This represents a simplified version of a model that in real life would use a notional space with more than three-dimensions to accommodate a greater number of variables.

Dr Flaxman says: "Every single person is at a point in this space. Some had defaults in their home loans and some didn't. You come along and we want to know what your classification should be. We draw a circle around you such that that circle contains, say, nine people" – Dr Flaxman draws a circle around nine dots – "and we take a majority vote. Among people like you, most had positive outcomes, so we make a positive choice."

"Does that explain to you why the decision about you was positive? In some sense. But I have to share the variables describing those other nine people with you. Am I allowed to do that? No. That's a privacy violation. And I usually no longer have that data."

While machines apply statistical techniques to identify dependencies between data on demographic variables and risks of default, Dr Flaxman's approach to explainability is to use statistical techniques to understand the machines. "We look for nonlinear dependencies," he explains, "not merely in the data that the machine is using, but between the data and the machine's predictions."

In addition to his research, Dr Flaxman has helped organise the Explainable Machine Learning Challenge, a competition run jointly by firms including Google and FICO and several universities that challenges teams of researchers to create machine learning algorithms that are both accurate and explainable.

HUMAN-LIKE COMPUTING

While work proceeds on making deep learning and other neural network approaches more explainable, other researchers at Imperial are advocating the renewed development of architectures that are more naturally suited to explainability.

Symbolic program synthesis is a longer established and quite different architecture. In this approach, a machine, learning from examples, generates a program that takes discrete symbols as inputs and performs computations over them to deliver an output. These algorithms, unlike neural networks, decompose into discrete symbols, and perform logical operations familiar from the computer programs written by humans. This means they are much more readily interpreted by humans.

Professor Stephen Muggleton

Stephen Muggleton, Professor of Machine Learning in Imperial's Department of Computing, endorses a view of AI first advanced by the researcher Donald Michie in the 1980s. According to this view, an ideal or 'ultra-strong' machine learning system would be able to use symbolic program synthesis to teach humans what it has learned, helping us improve our own performances.

Professor Muggleton laments the fact that present day implementations of machine learning have failed so far to adhere to this principle, despite their impressive abilities to carry out various sorts of task. He cites the example of DeepMind's AlphaGo, which was recently celebrated for being the first AI to beat a human champion at the ancient Chinese game of Go.

"There have been spectacular achievements by AlphaGo on beating Go players", Professor Muggleton says. "But after having made very interesting strategic wins, it would be desirable for the system to provide an explanation at the end that could advance the science of Go. This is an old science, written down and developed over years and volumes about how to play particular endings, and so on." Professor Muggleton smiles. "It seems a shame if we're not going to be able to carry on contributing to sciences like these because the humans are being cut out of what's being learnt."

Professor Muggleton recently collaborated with colleagues at the University of Bamberg in Germany to demonstrate that at least one state-of-the-art machine learning system is ultra-strong. The researchers found that experimental subjects struggled to perform a certain class of logical tasks on their own, but performed them more successfully after studying the logical hypotheses generated by the system.

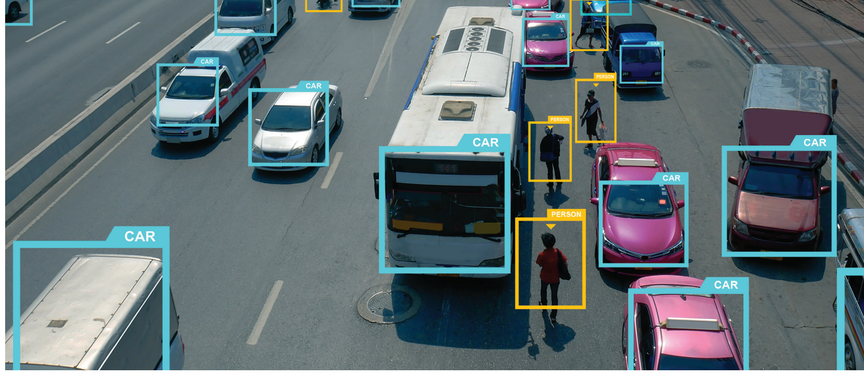
Explainability, in this sense, is at the heart of the new Human-Like Computing Network that Professor Muggleton is jointly leading. With funding from the Engineering and Physical Sciences Research Council (EPSRC), the network is bringing together 20 different UK research groups to develop systems that humans can learn from.

"The state of things at the moment is in some ways better than the 1980s. We have more powerful computers, and powerful learning techniques. But we're heading towards a future in which we progressively exclude humans from activities that they have excelled at in the past", he says. "There is an alternative to this, which is to have machine learning that is oriented towards comprehensible knowledge, where the activity of learning is seen as a joint activity between humans and machines. An international movement of research in this direction is emerging."

HYBRID APPROACHES

Besides lending itself naturally to explanation, symbolic program synthesis excels at learning from only a few examples, and applying the principles it learns reliably to a broad class of situations. However, because it works by

www.imperial.ac.uk/artificial-intelligence



manipulating discrete symbols, it traditionally requires discrete symbols as inputs.

"The problem is that the world isn't always crisp and symbolic", says Richard Evans, a Senior Research Scientist at DeepMind and PhD student in Imperial's Department of Computing. "Suppose you have a fuzzy image of a number and you don't quite know if it's a 4 or a 7." Situations like this are commonplace for machines that need to sense, such as robots or autonomous vehicles. Deep learning excels in these cases, because it can carry out the advanced statistical techniques required to classify noisy or mislabelled data.

Mr Evans is working with a colleague at DeepMind to develop a hybrid between symbolic and deep learning approaches. The approach uses a neural network architecture and is therefore good at classifying noisy sensory data. However, the network is designed to generate, through training, a program that represents the probabilities it is sensitive to symbolically. Because the symbols can be manipulated using universally-applicable logical rules, the system can apply the principles it learns to a broader class of situations than traditional deep learning can – and the programs it generates can be interpreted by humans.

A hybrid of symbolic AI and deep learning is also being developed under the auspices of the Human-Like Computing Network. "We have a collaboration going on with Nanjing University in China that's in the process of starting up what's called the Nanjing–Imperial Machine Learning Hub, where we're looking at techniques that combine statistical and symbolic machine learning", says Professor Muggleton.

Researchers have been trying to unify symbolic and machine learning approaches to machine learning for some years, but Mr Evans sees the hybrid approach as coming closer to fruition. "Recently, we have seen a handful of cases showing that this is indeed possible", he says. "These examples are proofs of concept and have scalability issues. But now, the challenge is less of a pure research problem, and more of an engineering and scaling problem."

ARGUMENTATION

Explainability means different things to different researchers. For some it means trusting AIs to perform properly; for others, learning from them. For some Imperial researchers, the idea is to develop systems with which we can engage jointly in rational deliberation.

Francesca Toni, Professor of Computational Logic in Imperial's Department of Computing, points to a series of posters on her office wall advertising past workshops and the inaugural lecture she gave in 2013 titled 'Could a machine ever argue?'.

Though 'argument' is a technical term in AI research, informally it means the same as it does in everyday life: "Not in the sense of clashing and fighting," Professor Toni explains, "but in a positive, dialectical and dialogical sense. You can ask me, 'Shall we go to watch that movie?', and I can say no, because I read some bad reviews. And you can say that a friend told you it's really cool. And so we are debating about whether or not to watch the movie."

When a machine argues – with itself, another machine or a person – it carries out a process similar to human deliberation. It takes one or more assertions (for example, my friend said that the movie is cool) and rules (if my friend said that the movie is cool then I should go to see it) and combines them to arrive at a reasoned conclusion.

One notable feature of this process is that it is accessible, firstly to expert users, who can view and understand the encoded argument the machine has generated, and secondly to end-users, as long as the AI is designed to represent its argument in a way that ordinary users can understand, for example natural language.

Professor Toni, a leading authority on argumentation, has been pursuing the approach for 20 years. "Right now, the field is hotter than it has ever been. It's hot because it is important and exciting. There is a big hype in AI, and a lot of the AI that is around is very mysterious" she says. "As humans we are quite used to arguing, in the positive sense of the word. So it seems natural to use argumentation as a methodology for constructing explainable AI."

www.imperial.ac.uk/artificial-intelligence