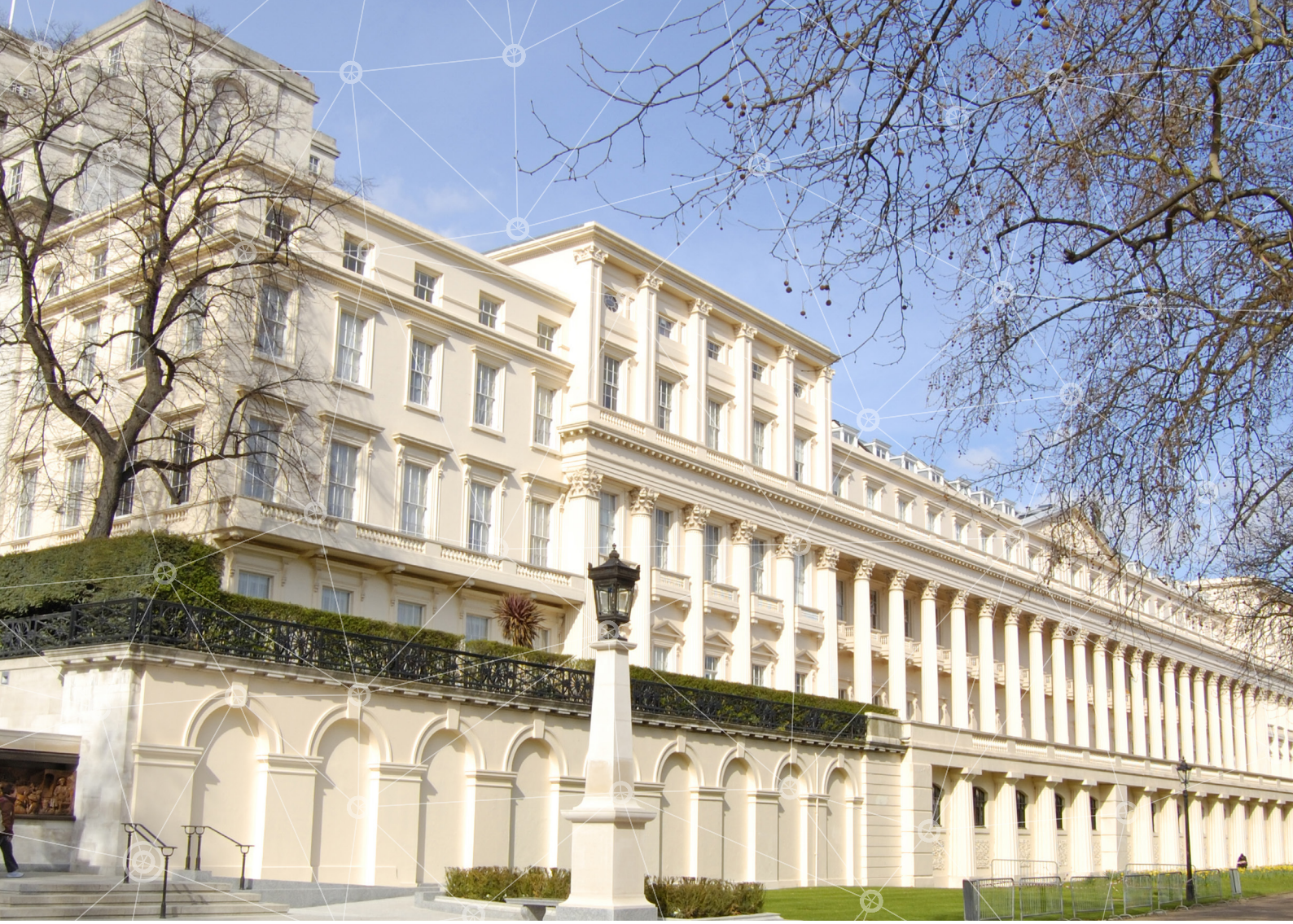
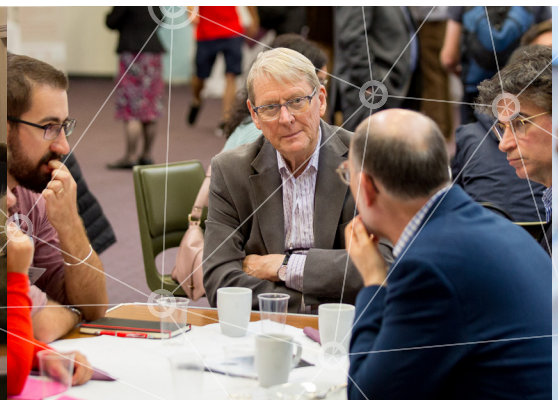


Validate AI

CONFERENCE WHITE PAPER

5th November 2019
THE ROYAL SOCIETY
6-9 CARLTON HOUSE TERRACE, LONDON SW1Y 5AG



Validate AI

Report from the conference held at the Royal Society On 5th November 2019

Contents:

Introduction

Automated decision making, via statistical, machine learning, and artificial intelligence (AI) algorithms has become central to our lives. They will become even more so in the future: the additional monetary benefit of AI to the UK's GDP has been estimated to be up to £232bn by 2030[1]. But this will only happen if we can be confident that the algorithms are fit-for-purpose, safe, reliable, and trustworthy, and that they will continue to be so as the world about them changes and develops.

The aim of the *Validate AI* conference was to explore how such systems can depart from this ideal, to examine tools and methods for ensuring sound and appropriate behaviour in a variety of different application domains under different circumstances, and to look at open challenges in validation. Issues explored included accurate and unbiased performance and its evaluation, model testing and formal verification, ensuring resilience against adversarial attacks, and the effective maintenance of systems as their working environment evolves. Representatives from public, private and academic organisations were brought together to give their perspectives on the issues and the way forward, and to share ideas and solutions.

The next section gives a summary of key points made at the conference. This is followed by a more detailed review of the proceedings.

Key points

AI research in the UK has in the past experienced two *AI Winters*, in which reaction to the slow progress in fulfilling inflated promises led to the drying up of research funding. While there are doubtless currently some inflated expectations, it is generally the case that the ubiquity of large data sets, coupled with the continuing streaming and accumulation of data (e.g. from transaction processing, administrative exercises, the web, the Internet of Things, etc.), and allied to the advent of powerful computers will mean that a third AI winter is unlikely. However, to minimise any potential backlash it is critically important that AI systems are properly validated, so that one can be confident that they fulfil their purpose. In short, we need to deliver AI systems that we can *trust*.

1. Consideration must be given to whether we should judge the performance of AI systems by a standard higher than that of natural systems (e.g. is one death caused by an autonomous vehicle worse than several or many caused by human drivers?). Public attention and reaction is likely to be focused on mistakes made by AI systems in contrast to humans (as is illustrated by the attention paid to autonomous vehicle crashes), and there is a need to tread carefully to avoid a backlash. The dangers are illustrated by the poor handling of the care data system[2], and the failure to communicate the balance of risks and benefits.

[1] https://www.pwc.co.uk//economic-services/assets/ai-uk-report-v2.pdf

[2] https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-015-2180-9

2. Given that AI systems are often intended to function in a human social environment, evaluation and validation of systems must take into account the social and ethical context. Ethics and trust in AI are very important considerations: we want AI systems that are fair, operate within our social and legal norms, and which respect people's privacy and do not misuse data. Validation of AI systems also need to cover these requirements. This will require input from professionals drawn from the wider society such as lawyers, philosophers and political leaders, not just AI technologists.

3. Validation should include the challenge of data from multiple sources. Heterogeneous (and sometimes contradictory) data sources should be included, since real data often have these characteristics. Validation in such contexts presents particular challenges.

4. The complete context in which an AI system will function must be taken into consideration when validating systems. In particular, there is the need to consider how systems might interact with other, possibly new and as yet unseen, processes.

5. In a similar vein, the entire life cycle of an AI system must be considered when undertaking validation. Almost by definition, AI systems will be functioning in changing environments so that, to the extent that it is possible, validation needs to be future-proofed against changing data characteristics as well as changing circumstances and environments.

6. At present we have only "weak" AI systems – systems which work in relatively well-defined domains with relatively well-defined objectives. The validation challenges posed by these are tough enough, but the validation challenges posed by "strong" AI systems (which are able to abstract and generalise to novel situations in the way human brains can) will pose much harder problems.

7. Validation of AI systems should include checks on how robust they are to small meaningless changes in the data or conditions under which they are operating. Indeed, recent research with deep learning systems has demonstrated they can be vulnerable to small changes. Adversarial attacks on these vulnerabilities could be used to sabotage an AI system. Validation systems need to be in place to detect these possible problems.

8. That there is often a tension between robustness and accuracy that needs to be considered when validating systems. This tension can arise when accurate high-performing systems are built for one particular data set (or over one particular time interval) but then fail to do well on other data sets (or later times). An example is given by credit risk models built on data preceding the 2008 financial crash, and then applied after the crash. Extra regularisation [3] can lead to more robust systems, but at the cost of local accuracy.

[3] Regularization is a technique in statistical modelling and machine learning to penalize for model complexity and hence reduce spurious overfitting to data of unnecessarily complex models.

9. An important aspect of AI systems validation is the system's capacity for explaining its output, or for having (the reasons for) its output explained. This needs to take account of who the explanation is for, since different respondents will require different levels of explanation. Explainability is often associated with superior generalisability and greater robustness because of the natural regularisation implicit in human understanding and mental modelling of phenomena.

10. The distinction between data-based models, derived purely from observed empirical relationships in data, and theory-based models, derived at least partly from an understanding of the processes involved has an important influence on validation. The former have a greater risk of brittleness [4] or fragility, especially when applied in nonstationary contexts.

11. It is important to get the fundamentals of data analysis right: to ensure that the analyst understands the data, the quality of the data, and the provenance of the data. The provision of more complex big data makes this difficult, but remains important: big data is not necessarily good data.

12. Simulation is an important tool for real-world applications, such as self-driving, when real errors can be expensive. Soft testing on simulated environments can prepare an AI system before it gets to the real world, and can also allow for simulation of extreme events to test the robustness of the AI, with minimal real-world costs.

[4] A system is brittle if it performs well on one specific task, but with a slight change to the task, performs badly.

Overview of proceedings

Lord Willetts began the conference by stressing the importance of the topic, suggesting that AI was now too deeply embedded in society for a third AI winter to occur. He drew attention to the carbon footprint of computer systems and the key challenge of how AI systems interact with humans. Ethical matters of data science, machine learning, and AI are attracting growing attention, with several bodies focusing on aspects of them having been established recently. The *explainability* of AI decisions was a key aspect of their acceptability, but we need to recognise that "artificial intelligence" is not the same as "natural intelligence".

David Hand presented eight dimensions of invalidity in AI systems, giving examples of each. The dimensions were: (1) Ignorance of the mechanisms of AI systems; (2) Ignorance of the limitations of AI systems; (3) Defining and ensuring the limits of behaviour; (4) Adequacy of data; (5) Robustness to perturbations of data; (6) Requiring assurance that AI systems do what they are supposed to do in *familiar* situations; (7) How AI systems behave with insufficiently specified problems; (8) How people use, or work with systems. He suggested that AI systems were confronting us with a new kind of principal agent problem, and that the law of unintended consequences was likely to manifest often. He consolidated the validation questions as (a) what properties must an AI system have for us to trust its decisions? (b) how can we ensure it has these properties? He also illustrated how the choice of evaluation method and performance measure can lead to dramatically different assessments of statistical models, highlighting the need for care when determining *how* to validate an AI system.

Marta Kwiatkowska observed that the data aspects discussed by Hand and the code aspects she would discuss were the two foundational aspects of AI. While the ambition was *strong* AI, currently at best we had only *weak* AI. She illustrated the challenges of validation with some less successful projects: IBM cancelling its Watson oncology project and Apple face-identification being defeated by a 3D printed face mask. She stressed the key role of trust in AI systems, drawing attention to the complex scenarios in which such systems would function, as well as the challenges of safety critical systems. Provable guarantees of performance were needed, distinguishing between verification, being proof the system satisfies its specification, and validation, being confirmation that it was fit for purpose. It is critically important to involve domain experts, since machine learning is very different from conventional programming. The former involves black box programming by pattern matching from examples, and it is important to ensure that systems are developed for users, not the developers themselves. Development must take place within the context of trust, ethics, morality, and social norms.

Michael Wooldridge focused on multiple interacting systems, which need to cooperate, coordinate, and negotiate with each other. He gave the example of financial trading systems, referring to the 2010 Flash Crash as an example of how hidden correlations between systems which behave in the same way can cause dramatic unexpected consequences and unstable behaviour. The ultimate goal in AI systems was not a system which has to be told what to do but one that works to help you. This means that systems need to know your preferences and values. This leads to the preference elicitation problem: how to code up our requirements for ethical and trustworthy behaviour in AI. He noted that

unstable equilibria were a particular challenge for multi-agent systems, and described the two strategies for understanding and tackling the issues: treat the problems (e.g. the Flash Crash) as a bug vs use agent-based modelling and simulation.

Aldo Faisal discussed diagnostic systems which mimic clinicians' perceptual ability to assess a situation – a very successful illustration of weak AI. He stressed the value of reinforcement learning, and of trust and explainability. This is partly a problem of human cognition: what is it we want to be explained and how? Do we want to interpret the AI or explain one specific decision made by an AI? One particular challenge arose from the regulators' requirement that systems should be fixed and not change and adapt (in unpredictable ways) as they autonomously learn.

Frankie Kay discussed the increasing use and challenges of combinations of data from multiple sources in government and the potential of efficiency and improved opportunities for applying AI. For example, in policing, we now have data from CCTV, biometrics and, increasingly, the Internet-of-Things. Dealing with this varied data requires that we ensure the quality of the data, that it is unbiased and fit for the purpose for which it is intended. It is important that industry and government using AI have staff with the technical skills to implement these systems correctly.

Shakeel Khan described his experience of AI capability building and in particular development (identify/adopt/innovate), knowledge advancement (explicit/tacit), and cross-sector collaboration (the triple helix framework). He stressed the importance of robustness across the entire life cycle and the need for proper validation management policies, drawing attention to *The Predictive Analytics Handbook* used at HMRC for development of predictive models. A particular challenge he identified was whether a system is fit for purpose as the population it is applied to changes; this is a common problem for systems that model human behaviour, such as credit scoring models. He suggested the use of sensitivity analysis or stress testing as a means to check for robustness to a changing population. Other issues identified were problems with future data loads, external economic or political factors, and the difficulties of predictive data not being available in the future.

Jasmine Grimsley looked at the use of AI in government, stressing AI project ethics in law (e.g. the GDPR), internal policies (e.g. ONS web-scraping policies), and advisory guidelines (e.g. EU ethical AI guidelines). Consideration had to be given to personal as well as social ethics.

Jonathan Crook described some of the challenges of validating AI systems in consumer finance, with focus on predicting default. Topics he discussed included sample distortion, unbalanced data, interpretability, and indirect bias in different groups leading to challenges to fairness.

Dan Kellett described the drift from relatively simple models in the consumer finance industry, such as logistic regression trees, to more sophisticated and less transparent models, such as gradient boosted methods. Validation questions he posed included: how does the model behave when presented with data values outside its previous range, is the model degrading over time, how well does the model perform on different products, is the model appropriate for a new population, and is the environment changing (e.g. economic

changes)? Sensitivity analysis is useful for this purpose - testing how the model performs when there are deviations in values of variables. He also pointed out the need to be able to answer the questions of how well we understand the data, how well we understand the model, and how will the model interact with downstream processes. He emphasized the need to properly understand the business (application) problem the model is intended to address and to ensure the appropriate use of models (i.e. a model built for one purpose may not be suitable for another).

Stan Boland discussed validation issues with self-driving cars. He suggested that self-driving cars are close to implementation but that currently their failure rates are too high (order of 10^{-3} failures/decision) relative to failure rates among human drivers (estimated as an order of 10^{-7} failures/decision). This is due to a very long tail of perception problems. Therefore, there is a great deal of room for improvement. Noting that it was not possible to model or simulate all possible scenarios an autonomous vehicle might encounter, he drew attention to the need to model the relationship between physical invariants (such as rain) and the error rate of perception systems, in an attempt to alleviate brittleness risks. He proposed a framework for this based on simulating perception systems through generating all salient aspects, adding noise to images and deliberately simulating adversarial examples to improve robustness, generating new directed tests, measuring performance, and establishing metrics. Simulation is a large part of the development and testing process for self-driving cars. Skills and experience of programmers from the gaming industry are employed to ensure the development of realistic driving simulations. For validation, standard machine learning evaluation can be supplemented by domain knowledge to ensure the system is behaving appropriately by adding constraints, such as the expected behaviour of traffic lights (changing through red, amber, green), for example.

Iain Whiteside noted that autonomous vehicles might well fail to recognise unusual dangerous situations. He gave the illustration of a cyclist, wearing headphones in the dark and with a large mirror strapped to her back, and commented that “safety is heterogeneous”. Taking a deep perspective, he noted that the code for neural networks is in a sense simple, involving matrix multiplication: the “bugs” are really problems with the data. He noted the tension between accuracy vs the time and data needed to train systems. Iain also recommended testing autonomous vehicles using scenarios, or safety cases, drawing attention to the valuable resource for researchers in AI Validity given at https://nsc.nasa.gov/resources/case-studies produced by NASA, which contains case studies of how systems fail.

Michael Bronstein pointed out that the success of deep learning was a consequence of the abandonment of the universal approximation, and its replacement by the imposition of translational invariance. He stressed the adaptation of this idea to graphs through local permutation invariance. He used the Netflix recommender system to illustrate graph theoretical ideas, looking at various challenges, such as how to update as new data comes in.

Pushmeet Kohli opened by asking why it is that image classifiers can perform excellently on one image data set, but fail miserably on a different one: with this question in mind, he discussed the need for rigorous training, robust AI, and the verification of AI systems, stressing the need for insensitivity to meaningless changes in the data, and noting how

slight changes to the system can dramatically reduce performance. He described strategies for tackling such problems, for example using adversarial systems. He commented that an approach to randomly generate examples to improve robustness does not always work because this approach may not capture all sensitive scenarios. To cope with this, he and his colleagues are working on mathematical approaches based on buffer zones of neighbourhoods around points to be classified. Pushmeet stressed the need for rigorous testing, adding that AI software needs to conform to social norms for safety and values, through the formal specification of robustness, fairness and compliance. He discussed "explainability" and the need to consider *who* requires the explanation and *what* needs to be explained. He distinguished between explaining a learnt system and learning an explained system, and noted that the natural inductive biases in human language means that explainable models can generalise better than unexplainable models – provided those biases regularise in the right way. He also noted the fundamental point that the challenges were not merely technical, but were also social.

Giles Herdale described the so-called *Peelian Principles*, named after Sir Robert Peel, and in particular the idea that "the police are the public and the public are the police". He discussed drivers for data-driven policing, and the challenge posed by the explosion in digital evidence and changing offending patterns. In the context of these challenges, he discussed independent oversight, internal management and expertise, and public engagement. He also drew attention to the critical questions of what the public are seeking to achieve and of how to measure success.

Stephanie Hare commented on the evolution from first generation biometrics (e.g. fingerprint, DNA) to second generation (e.g. face, eye, gait recognition), stressing the critical importance of joining up data, and noting that the police need a legal framework for the use of technology. In particular, she highlighted issues of privacy and misuse of data in AI systems. In many parts of the world, including UK and other western countries, second generation biometrics can be taken without consent or even knowledge: there is currently no legal protection. Many police forces and legislatures across the world are now reviewing this situation.

Panel Session. The panel discussed the explosion of interest in AI ethics and the centrality of ethics in AI development. The nature of ethical failures in the AI world was contrasted with that in (for example) engineering. It was suggested that further discussion of ethics in AI was not required – we know what we need to do – what is required is development of appropriate methodology to ensure the ethical requirements are implemented and validated correctly. An analogy was drawn with clinical trials in medicine which is strongly regulated: perhaps a similar approach is required for AI? The question was raised about whether a machine learning system can be considered flawed because it encodes discrimination and sociodemographic inequalities found in society. There was a general consensus that, no, it couldn't, but we have the opportunity to build machine learning systems that support rectifying the problems of discrimination in society. Care is needed when dealing with this question since discrimination in machine learning may in fact be a consequence of bias in the training data. Other topics discussed by the panel included: the extent to which we should seek to encode emotions in AI systems, the fact that much knowledge is tacit and data-based (e.g. how anaesthetics work), the use of AI systems in accountancy, the perennial issue of models encapsulating data bias, whether we should

hold AI systems to a higher level of accountability than humans, the fact that “data science” is not (yet) a profession, the suggestion that most ethical breaches occur because people are not properly trained, the challenge of breaking down barriers between technical and policy communication, and the question of how things will change in the future, with the practical comment that maybe we have to accept the sub-optimal for now, knowing it will improve in time.

Next Steps

We greatly appreciate the contribution of all those who attended the Validate AI conference and in particular to the presenters who made the event such an informative day.

We are now in consultation with existing and prospective supporters of the Validate AI Conference and are considering the best way forward, including a follow-on conference in 2020.

We look forward to ongoing engagement in relation to our objectives to:

1. Promote targeted academic research on topics raised by the Validate AI conference in relation to technical and ethical considerations in the deployment and maintenance of such systems.

2. Encourage government-supported initiatives to promote collaboration between public, private and academic sectors. Knowledge Transfer Partnership, run by Innovate UK, is a great example of cross-sector collaborative working.

3. Contribute to forums to promote dialogue across the Validate AI community.

4. Reach out to partners globally to champion and refine the thinking of the Validate AI movement.

We would be delighted with any suggestions you may have to progress the Validate AI initiative. You can e-mail us at contact@validateai.org

We will report back our next steps in the first quarter of 2020.

Yours Sincerely

Validate AI Organising Committee

Appendix 1: Programme

08:45-09:15 Registration

09:15-09:17 Welcome: David Hand

09:17-09:25 Opening remarks: Lord Willetts

09:25-09:55 Why is AI validity and maintenance critical? : David Hand and Marta Kwiatkowska

09:55-10:25 Understanding the dynamics of multi-agent systems: Michael Wooldridge

10:25-10:55 Diagnosis and prognosis systems in health: Aldo Faisal and Yike Guo

10:55-11:25 Adoption and sustainability of machine learning maintenance – lessons from government: Shakeel Khan, Frankie Kay, and Jasmine Grimsley

11:55-12:55 Financial sector: Dan Kellett and Jonathan Crook

12:25-13:25 Lunch

13:25-13:55 Challenges of assuring safe self-driving: Stan Boland and Iain Whiteside

13:55-14:25 The promises and challenges of machine learning on graphs: Michael Bronstein

14:25-14:55 Towards robust and explainable artificial intelligence: Pushmeet Kohli

14:55-15:25 Law and order: Stephanie Hare and Giles Herdale

15:55-16:55 Panel session: Zeynep Engin, Carly Kind, Marta Kwiatkowska, Martin Goodson, and Tom Smith

16:55-17:00 Closing remarks: David Hand

A display of posters showcasing student work in the area was also presented. Joseph Brook won the prize for best poster.

Appendix 2: Organising Committee

Dr Zeynep Engin – University College London
Professor Yi-Ke Guo – Imperial College London
Tom Smith – ONS Data Science Campus
Dr Tony Bellotti – Imperial College London (Conference Proposer)
Professor David Hand – Imperial College London (Conference Proposer)
Professor Marta Kwiatkowska – Oxford University (Conference Proposer)
Shakeel Khan – Her Majesty's Revenue and Customs (Conference Chair & Proposer)

Appendix 3: Sponsors and Supporters

Our Sponsors

Imperial College London

Ada Lovelace Institute

THE OPERATIONAL RESEARCH SOCIETY

Our Supporters

UCL

UNIVERSITY OF OXFORD

The Alan Turing Institute

Data Science Campus

GORS Government Operational Research Service

DATA FOR POLICY .ORG

CapitalOne

Office for National Statistics

CAMBRIDGE UNIVERSITY PRESS

SGAI

UNIVERSITY OF EDINBURGH Business School

conteur

Independent DIGITAL ETHICS Panel for Policing

EvolutionAI

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS

FIVE AI

NICE National Institute for Health and Care Excellence

University of Nottingham UK | CHINA | MALAYSIA

DATA & POLICY

Contact us at: contact@validateai.org

Follow us on twitter: @Validate_AI, #ValidateAI

Visit our website: validateaiconference.com

Linkedin: https://www.linkedin.com/company/validateai/