Proceedings of the Oxford Computer Science Conference 2026

General Chairs: Jayroop Ramesh and Jumana Baghabrah
Conference Committee: Hunar Batra and Allison Clement
With assistance from: Sarah Retz-Jones and Jodie Mattioli

24th June 2026



Contents

1 Programme **2**

2 OxWoCS Keynote Speaker **3**

3 Abstracts of Talks **4**

4 Abstracts of Lightning Talks **8**

5 Abstracts of Posters **10**

Programme

09:30	Registration and Coffee		
10:00	Introduction by DGS and Deputy Agenda Walkthrough	Jonathan Barrett CoGS Committee	
10:30	Oral Session I: Computational Theory	Hera Brown	Decidability of Extensions of Presburger Arithmetic by Hardy Field Functions
		Maximilian Frank	The Trade-Off Triangle of Universality, Consistency and Totality in Semantic Mappings, and Why Natural Language Models Cannot Guarantee Consistency
11:00	Coffee Break		
11:15	Lightning Session I: AI Applications	Hunar Batra	*MMDiff: Multimodal Model Diffing for Feature Discovery and Control*
		Owen Green	Effective Altruism is Hard: Strategic Voting in Multiwinner Elections
11:30	Oral Session II: Programming & Verification	Haoxuan Yin	Contextual MetaML: Syntax and Full Abstraction
		Syyeda Fatmi	Towards Robust Behavioural Equivalence
12:00	Lunch and Posters	Allison Clement	Confidence in Angle Predictions for Clinical Decision Support
		Benjamin Hardin	What Can Eye Gaze Teach Us About Real-World Cycling? Insights From the Oxford RobotCycle Project
		Sophie Fischer	Navigating Clinical Uncertainty in Medical Segmentation
		Elena Lickel	Inverse Reinforcement Learning for Firm Behaviour in Macroeconomic Agent-Based Models
		Filip Jasionowicz	Finding Induced Detours in Linear Time on Interval and Circular-Arc Graphs
		Merlin Plant	Agentic Data Cards: A Metadata Schema for Agent-to-Agent Data Transmission
		Rachel Scanlon	What can Bodleian Libraries do for me?
14:00	OxWoCS Keynote Speaker	Dr. Mengyun Qiao	From Models to Discovery: Generative and Agentic AI for Cardiovascular Research
15:00	Lightning Session II: Modelling	Rohan Acharya	Recent Developments in Randomised Time-Bounded Kolmogorov Complexity
		Guolei Zeng	Normality Calibration in Semi-supervised Graph Anomaly Detection
15:15	Coffee Break		
15:30	Oral Session III: Algorithms	Sichao Xiong	A Parameter-Free First-Order Algorithm for Non-Convex Optimization with $O(\epsilon^{-5/3})$ Global Rate
		Mahyar Afshinmehr	EFX Allocations Exist for Multigraphs
16:00	Oral Session IV: AI and Society	Daniel Marks	Pick-to-Learn: Meta-learning State-of-the-art Generalisation Guarantees
		Jack Harris	Striking the Balance: AI, Communal Resources, and Community in Computer Science
16:30	Panel Deliberation/Voting		
17:00	Prizes		
17:30	End of conference. Walk over to Reuben College		
18:00	Drinks reception at Reuben College		
18:30	Conference dinner at Reuben College		

OxWoCS Keynote Speaker

This year's Keynote Speaker was kindly organised in conjunction with Oxford Women in Computer Science.

From Models to Discovery: Generative and Agentic AI for Cardiovascular Research

Abstract

Cardiovascular imaging studies now provide detailed measurements of cardiac structure and function at unprecedented scale. While these data have enabled increasingly accurate predictive models, an important challenge remains: how can we translate large collections of imaging and clinical data into meaningful biological and clinical discoveries?

This talk presents a research journey from population-scale cardiac modelling to AI systems designed to support scientific discovery. It begins with multimodal integration and four-dimensional representations of the heart, followed by generative models that learn the distribution of normal cardiac anatomy and motion from large population cohorts. These models provide personalised reference spaces that enable the identification and quantification of patient-specific deviations from expected cardiac patterns.

Building on these representations, the talk explores how agentic AI systems can support downstream tasks such as phenotype construction, hypothesis generation, association analysis, and clinical interpretation. Examples from cardiac MRI and population imaging studies will illustrate how generative and agentic approaches can help connect large-scale data with clinically relevant questions.

Biography

Dr. Mengyun Qiao is a Lecturer in the Department of Mechanical Engineering at University College London (UCL). Her research focuses on artificial intelligence for cardiovascular imaging, with interests in generative modelling, multimodal learning, and agentic AI systems for clinical discovery. She leads the Learning and Understanding in Medical AI (LUMA) Lab at UCL and develops computational approaches for analysing cardiac structure, function, and disease at population scale. Her work has been published in leading venues including Nature Machine Intelligence, Nature Cardiovascular Research, IEEE Transactions on Medical Imaging, and MICCAI.

Abstracts of Talks

Oral Session I: Computational Theory

Decidability of Extensions of Presburger Arithmetic by Hardy Field Functions

Hera Brown

We study the extension of Presburger arithmetic by the class of sub-polynomial Hardy field functions, and show the majority of these extensions to be undecidable. More precisely, we show that the theory $\mathrm{Th}(\mathbb{Z}; <, +, \lfloor f \rceil)$, where f is a Hardy field function and $\lfloor \cdot \rceil$ the nearest integer operator, is undecidable when f grows polynomially faster than x. Further, we show that when f grows sub-linearly quickly, but still as fast as some polynomial, the theory $\mathrm{Th}(\mathbb{Z}; <, +, \lfloor f \rceil)$ is undecidable.

The Trade-Off Triangle of Universality, Consistency and Totality in Semantic Mappings, and Why Natural Language Models Cannot Guarantee Consistency

Maximilian Frank

Communication and interoperability can become difficult if two systems use different terms for the same meanings. This is especially true for model checking if the model and the verification rules use different semantics. In such a case, the model checker could not connect the concepts in each. To bridge the gap between different semantics, one requires a semantic mapping (or ontology mapping) from one to the other.

We establish that such a semantic mapping cannot be consistent (i.e. free from contradictions), total (i.e. mapping all concepts) and universal (i.e. applicable to all source semantics in a domain) at the same time, if the source semantics can be arbitrary and thus there exist at least two possible semantics that contradict each other.

This has multiple implications. For example, natural language models cannot be perfectly consistent. Similarly, privacy compliance checking for arbitrary software may trade off universality, if it would have to maintain consistency and totality (i.e. assessing the whole system). The trade-off triangle also shows that – in theory – there exists a single mapping between all non-contradictory standards, which would therefore all make them interoperable.

Oral Session II: Programming & Verification

Contextual MetaML: Syntax and Full Abstraction

Haoxuan Yin

MetaML-style metaprogramming languages allow programmers to construct, manipulate and run code. In the presence of higher-order references for code, ensuring type safety is challenging, as free variables can escape their binders. In this paper, we present Contextual MetaML, *the first metaprogramming language that supports storing and running open code under a strong type safety guarantee*. The type system utilises contextual modal types to track and reason about free variables in code explicitly.

A crucial concern in metaprogramming-based program optimisations is whether the optimised program preserves the meaning of the original program. Addressing this question requires a notion of program equivalence and techniques to reason about it. In this paper, we provide a semantic model that captures contextual equivalence for Contextual MetaML, establishing *the first full abstraction result for an imperative MetaML-style language*. Our model is based on traces derived via operational game semantics, where the meaning of a program is modelled by its possible interactions with the environment. We also establish a novel closed instances of use theorem that accounts for both call-by-value and call-by-name closing substitutions.

Towards Robust Behavioural Equivalence

Syyeda Fatmi

Model checking is a formal verification technique used to automatically determine whether a hardware or software system satisfies a desired property by analysing a mathematical model of its behaviour. For probabilistic systems, these models incorporate uncertainty in system behaviour, but their complexity often necessitates simplification through state-space reduction techniques such as probabilistic bisimilarity, which merges behaviourally equivalent states. However, probabilistic bisimilarity is not always robust with respect to the transition probabilities, which are often approximated in practice. As a result, states deemed equivalent may exhibit significantly different behaviour under slight changes, potentially compromising verification results. To address this issue, we introduce robust probabilistic bisimilarity, a new equivalence notion that preserves approximate behavioural equivalence under small transition probability perturbations. This contributes to the broader goal of making formal verification more resilient to modelling inaccuracies in increasingly complex and data-driven systems.

Oral Session III: Algorithms

A Parameter-Free First-Order Algorithm for Non-Convex Optimization with $O(\epsilon^{-5/3})$ Global Rate

Sichao Xiong

We introduce PF-AGD, the first parameter-free, deterministic, accelerated first-order method to achieve an $O(\epsilon^{-5/3})$ oracle complexity bound when minimizing sufficiently smooth, non-convex functions; this is the best-known bound for first-order methods on smooth non-convex objectives. Unlike existing methods possessing this rate that require a priori knowledge of smoothness constants, we use an adaptive backtracking scheme and a gradient-based restart mechanism to estimate local curvature. This yields a practical algorithm that matches best-known theoretical rates. Empirically, PF-AGD outperforms the practical variant of AGD-Until-Guilty (Carmon et al.), as well as other parameter-free variants, and is a viable alternative to nonlinear conjugate gradient methods.

EFX Allocations Exist for Multigraphs

Mahyar Afshinmehr

We study the fair allocation of indivisible goods among agents, with a focus on limiting envy. A central fairness notion is envy-freeness up to any good (EFX), which requires that any envy toward another agent vanishes after the removal of any single good from the latter's bundle. The existence of EFX allocations is considered a major open problem in fair division, and has so far been established only in limited settings.

Christodoulou et al. [2023] proved the existence of EFX allocations for graphical valuations. In this setting, agents correspond to the nodes of an underlying graph, goods correspond to edges, and each good has positive value only for the endpoints of the corresponding edge. Their proof crucially relies on the restriction that the graph is simple, meaning that for any pair of agents, there is at most one good that has value for both agents.

For multigraph valuations, where multiple goods may be valued by the same pair of agents, only partial results are known. Amanatidis et al. [2024] and Kaviani et al. [2025] obtained 2/3 and $\sqrt{2}/2$ approximations of EFX, respectively; Kaviani et al. [2024] established existence under restricted additive valuations; and Afshinmehr et al. [2025a] proved existence under the assumption that the shortest cycle containing non-parallel edges has length at least 4.

In this paper, we resolve this open problem by proving the existence of EFX allocations for multigraph instances under cancelable valuations, a strict superclass of additive valuation functions. Our proof is algorithmic and computes such allocations in polynomial time when the valuation functions are additive, and in pseudo-polynomial time when they are cancelable. This work contributes to the small number of EFX existence results that apply to an arbitrary number of agents.

Oral Session IV: AI and Society

Pick-to-Learn: Meta-learning State-of-the-art Generalisation Guarantees

Daniel Marks

Data-driven methods are increasingly being deployed in safety-critical settings, including autonomous driving, robotics, and industrial automation. In such domains, a central challenge for practitioners is to assess whether a learned model will make reliable predictions on data it has not seen before. Existing approaches, including conformal prediction, test-set methods, and PAC-Bayes bounds, face an important trade-off: they either require part of the dataset to be set aside in order to produce guarantees, potentially degrading the quality of the learned model, or they yield bounds that are often too loose to reflect true performance. We propose a general procedure, called Pick-to-Learn, that avoids this trade-off by allowing any learning algorithm, such as a neural network, to use all available data while still producing a formal certificate on the probability of making a mistake. Pick-to-Learn induces a sample compression scheme from any learning algorithm, thereby enabling tight compression bounds to be applied to models that may not natively exhibit any compression structure. Numerical experiments across machine learning and data-driven control tasks show that Pick-to-Learn can improve the state of the art in both post-training performance and certified guarantees.

Striking the Balance: AI, Communal Resources, and Community in Computer Science

Jack Harris

Whatever happened to Stack Overflow? AI has raced ahead and provided a variety of new and powerful tools for software development, but little attention is paid to what has happened to the communal resources that predicated the development of Large Language Models. This talk will explore the communal resources and communities that computer scientists created to make software engineering easier, such as Stack Overflow and the open source software movement, and examine the ways in which they are currently being degraded by AI. Economic concepts, such as the Tragedy of the Commons, will then be applied to these examples, both to better understand the problems faced, and to explore potential solutions.

Abstracts of Lightning Talks

Lightning Session I: AI Applications

MMDiff: Multimodal Model Diffing for Feature Discovery and Control

Hunar Batra

Multimodal Large Language Models (MLLMs) exhibit strong visual understanding, yet the internal features that cause these behaviors remain difficult to identify, audit, or control. While applicable to post-hoc inspection, hidden states that are decomposed into interpretable feature directions using sparse autoencoders (SAEs) do neither readily isolate which features are changed by multimodal training, nor are they directly useful for targeted control. We introduce MMDiff, a multimodal model-diffing framework that trains multimodal SAEs and turns them into feature-level interfaces for discovering and controlling multimodal behavior. MMDiff supports three uses: (i) feature isolation, by diffing a base-LM SAE against its multimodal-adapted counterpart to identify features altered by multimodal training; (ii) task-specific feature detection, via per-token contrastive firing analysis that isolates causal features; and (iii) feature-level control, by causally removing or steering the discovered feature directions. We train multimodal SAEs for two MLLM families, LLaVA-MORE and PaliGemma 2, and evaluate on visual-spatial understanding, multimodal safety, and OCR. MMDiff discovers sparse, causally specific features whose removal selectively degrades target behaviors by an average of 12% on spatial tasks and 17% on OCR, and reduces attack success rate by 24% on multimodal safety attacks, with no impact on VQA performance. Steering these features improves spatial and OCR accuracy by +3.6% and +1.5% on average over a standard single-layer steering baseline. These results show that multimodal SAEs can serve not only as interpretability tools, but as mechanisms for auditing, steering, and controlling MLLMs behavior toward safer and more capable generations.

Effective Altruism is Hard: Strategic Voting in Multiwinner Elections

Owen Green

Approval-based committee (ABC) elections involve selecting a committee of size k from a set of m candidates using the preferences of n voters. The study of the strategic behaviour of voters has received little attention for this type of election. We propose a new model for studying this behaviour based on the notion of multiselves equilibria which reflects the idea that voters who have similar preferences over candidates should vote in similar ways. In this model, we establish several computational hardness results concerning the evaluation and optimisation of the strategies

of voters in ABC elections even when all voters are altruistic. In this talk, I will briefly describe the model before giving an overview of one NP-hardness result which utilises a game theoretic perspective on the Motzkin–Straus theorem.

Lightning Session II: Modelling

Recent Developments in Randomised Time-Bounded Kolmogorov Complexity

Rohan Acharya

We motivate some definitions of meta-computational problems based on notions of Kolmogorov complexity. Furthermore, we explain some recent unconditional lower bounds against such meta-computational problems and some of the challenges to overcome in future work.

Normality Calibration in Semi-supervised Graph Anomaly Detection

Guolei Zeng

Semi-supervised graph anomaly detection (GAD), which assumes a subset of labeled normal nodes for training, is widely studied. However, existing methods learn normality only from these labeled nodes, often overfitting their patterns and causing high detection errors, such as many false positives. To overcome this limitation, we propose GraphNC, a graph normality calibration framework that leverages both labeled and unlabeled data to calibrate the normality from a teacher (a pre-trained semi-supervised GAD model) jointly in anomaly score and representation spaces. GraphNC includes two main components, anomaly score distribution alignment (ScoreDA) and perturbation-based normality regularization (NormReg). ScoreDA optimizes our model's anomaly scores by aligning them with the teacher's score distribution. Because the teacher provides accurate scores for most normal nodes and some anomalies, this alignment pulls the scores of the two classes toward opposite ends, making them more separable. To reduce the impact of inaccurate teacher scores, NormReg regularizes normality in the representation space, making normal node representations more compact via a perturbation-guided consistency loss applied only to the labeled nodes. Comprehensive experiments on six benchmarks demonstrate that GraphNC (1) consistently and substantially enhances the performance of teacher models from different GAD methods, and (2) achieves new state-of-the-art performance. Code is available at `https://github.com/mala-lab/GraphNC`.

Abstracts of Posters

Confidence in Angle Predictions for Clinical Decision Support

Allison Clement

Anatomical landmarks are used for clinical measurements, screening, and to guide treatment decisions. In this work, we explore the clinical application of landmark-based angle measurements, with a particular aim of screening infants for Developmental Dysplasia of the Hip.

Our automated machine method uses a simple UNet++ architecture. The network is used to predict landmark heatmaps, which represent landmark localisation certainty. A Monte Carlo-like approach is then used to approximate an angle distribution from landmark heatmaps. We propose a confidence metric from the derived angle distributions.

Multiple clinician annotations are combined and compared to the machine predictions. The machine-generated angle distribution is verified by confirming the correlation of the mean angle values and standard deviations per scan, between the multiple clinicians and the machine. The confidence scores correlate for the clinicians combined and the machine. The confidence of the machine strongly correlates with the sum of the confidence scores given by clinicians for each scan.

This work is the first to present a method for estimating the distribution of clinically relevant angles from predicted landmarks. Landmark-based angle confidence can establish robust methods and increase clinician trust in using automated or computer-aided methods.

What Can Eye Gaze Teach Us About Real-World Cycling? Insights From the Oxford RobotCycle Project

Benjamin Hardin

Although much is known about the physical danger of cycling situations, less is understood about the perceived danger of cycling. Furthermore, perception of danger may be filtered at a subconscious level and therefore difficult for one to self-report. To this end, these subconscious perceptions can be revealed through physiological metrics such as eye gaze. This paper explores the perceived safety of cycling in Oxford, United Kingdom and explores the ability of wearable eye tracking glasses to produce insights about the differences in perception under different environments and events. This paper finds that eye gaze patterns change between using bike lanes, car lanes and shared bus lanes, representing different cognitive challenges of each lane type. This paper presents that different intersections have significantly different eye gaze patterns which may have implications for cyclist stress. Finally, eye gaze patterns differ in the presence of events such as passes and pedestrians in

the road compared to when cycling with no events. This paper draws conclusions on the benefits and limitations of using wearable eye trackers to estimate stress and cyclist workload.

Navigating Clinical Uncertainty in Medical Segmentation

Sophie Fischer

Clinical uncertainty is an unavoidable issue in every area of medical imaging, and must be carefully considered in order to ensure trustworthy AI: errors that no clinician would ever make should be judged much more seriously than those in clinically disputed areas. Clinician uncertainty is also a multi-factored problem, including both the intra-uncertainty from a single annotator and the inter-uncertainty between annotators, and has many causes ranging from the quality of the scan to the annotation tool used. Modelling uncertainty is therefore a similarly complex task that varies across every task in medical imaging.

While many studies have focused on predicting AI uncertainty, understanding clinical uncertainty suffers from the lack of publicly available studies on intra- and inter-annotator variability. Our work considers how inter-annotator uncertainty is estimated, presented, and used for the task of medical segmentation, and the risks for AI of neglecting uncertainty. We focus in particular on how the evaluation metrics commonly used are poorly equipped to handle annotator uncertainty, and propose an alternative approach for future research.

Inverse Reinforcement Learning for Firm Behaviour in Macroeconomic Agent-Based Models

Elena Lickel

Macroeconomic agent-based models specify aggregate dynamics through the interaction of heterogeneous agents, but the behaviour of those agents is often implemented through hand-crafted adaptive rules. In these models, the modeller therefore designs the policy rules that agents follow, while the objectives underlying those rules often remain implicit. Replacing these rules with reinforcement learning shifts the modelling problem from policy design to reward design. This paper asks whether adversarial inverse reinforcement learning (AIRL) can address this challenge by recovering a reward function that can be reused as behavioural components inside a macroeconomic agent-based model.

The focus is on firms because they provide a credible benchmark for comparison with existing reinforcement-learning work in macroeconomic agent-based models, and because firm-level accounting data are available for the real-data transfer exercise. The question is evaluated in an agent-based model of the Italian economy using both synthetic and real-world data.

The synthetic experiments use simulator-generated demonstrations from reinforcement-learning firms with a known coded reward, providing a controlled benchmark for reward recovery. The real-data experiments use Orbis firm-accounting transitions mapped into the simulator, testing whether a reward inferred from observed firm behaviour can improve downstream macroeconomic simulation. In both settings, AIRL is trained on firm-level transitions. The recovered reward is then used to train firms inside the macroeconomic agent-based model, and the resulting simulations are compared with observed macroeconomic outcomes.

The synthetic results show that AIRL recovers rewards that support near-expert firm performance when used to optimise new reinforcement-learning firms inside the agent-based model. The Orbis

transfer results are more selective: the resulting simulations improve on several macroeconomic outcomes, including GDP, but do not dominate the calibrated heuristic across all variables.

The paper therefore shows that inverse reward learning can provide reusable behavioural components for macroeconomic simulation. Learned rewards do not replace calibration or economic judgement, but they offer a viable alternative to fully hand-coded firm objectives on targeted intervention margins.

Finding Induced Detours in Linear Time on Interval and Circular-Arc Graphs

Filip Jasionowicz

The problem of finding shortest paths in graphs is well explored. Meanwhile, finding detours, i.e. non-shortest paths, is an active area of research. In this work, we provide $O(n + m)$-time algorithms for finding induced non-shortest paths on the classes of interval and circular-arc graphs with n vertices and m edges. As to the state of the art, for interval graphs, a reduction to Ishizeki, Otachi, and Yamazaki's algorithm for finding longest induced paths in graphs of bounded chordality solves the problem in time $O(nm)$. For circular-arc graphs, the literature does not mention any algorithm improving on Chiu and Lu's $O(n^{4.75})$-time one for general graphs.

Agentic Data Cards: A Metadata Schema for Agent-to-Agent Data Transmission

Merlin Plant

Autonomous software agents exchange datasets across multi-agent pipelines, yet no existing standard supplies the machine-readable quality, fitness, and trust metadata they need at the point of transmission. This paper presents the Agentic Data Card (ADC), a three-layer specification that lets a receiving agent compare, validate, and select competing dataset offers. Developed through a Design Science Research synthesis of dataset documentation frameworks, agent communication protocols, and metadata catalogue standards, the artefact is evaluated through a walkthrough of a data procurement scenario in the text modality. It delivers three capabilities, one per layer: machine-readable structural comparison through typed schema fields and Design-by-Contract constraint assertions; attributed qualitative reasoning through the Attributed Text Block, a novel construct wrapping every narrative field with authorship, calibrated confidence, and review status; and trust-calibrated decision support through quality scoring, fitness-for-purpose declarations, and verifiable attestation chains. ADCs travel with datasets as JSON files through existing agent protocols, so a receiving agent can rank candidates, reject offers below a fitness threshold, or request iterative refinement. The Agentic Data Card is a prototype specification that unifies structural, qualitative, and trust metadata into a single artefact designed for agent-to-agent transmission.

What can Bodleian Libraries do for me?

Rachel Scanlon

Poster representing the Radcliffe Science Library, Bodleian Libraries, University of Oxford.