Table of Contents

Multi-source LLM-assisted clinical codelist construction with auditable human review: a responsible-AI case study

Track: Emerging Research

Authors: Carlos J Ramirez Suarez (University of Cambridge)

Abstract: Codelist construction for electronic health record research currently requires manual code assembly across vocabularies (Watson et al. 2017, Stage 2). clinicalcodes.uk is an open-source pipeline (Apache 2.0) addressing this step with multi-source candidate retrieval (OMOPHub, OpenCodelists, QOF Business Rules, and a ChromaDB semantic index over PubMedBERT embeddings), UMLS synonym enrichment, per-code scoring with Claude Haiku 4.5 at temperature zero, and a clinician review gate. Every output code carries source provenance, an LLM rationale, and a confidence score. Approved codelists carry a SHA-256 content-hash over the canonical-ordered final decisions, so post-approval edits to any persisted decision invalidate the signature. The pipeline is mapped against DCB0129 and DCB0160 clinical-safety standards in CLINICAL_SAFETY.md within the repository. Benchmarked against 15 published OpenCodelists references (13 SNOMED CT, 2 ICD-10), strict-view F1 improved from 0.49 to 0.57 after four targeted methodology fixes, with 95% BCa bootstrap CI [0.44, 0.68]. The pre-vs-post lift is significant under a paired McNemar's test on per-code correctness: chi-squared = 42.93, p = 5.7e-11 (n = 2,769 paired observations). A per-retriever ablation decomposes the marginal contribution of each retriever and of the LLM scoring step. The work has clear limitations: n = 15 is pilot-tier, reference labels were interpreted by a single non-clinician evaluator, and an inter-rater agreement study with multiple clinicians is required under any funded follow-on or NHS Trust pilot. Live demo: clinicalcodes.uk. Repository: github.com/carlos-ramblox/nice-clinical-codes.

From Principles to Practice: Embedding Ethics in an AI-Enabled Drone System to Support Battlefield Casualty Triage

Track: World Class Research Track

Authors: Peter Lee (University of Portsmouth)

Presenter: Peter Lee (University of Portsmouth)

Abstract: Much contemporary debate on military artificial intelligence focuses on autonomous weapons, yet AI enabled support systems designed to save lives raise equally complex ethical challenges. This paper presents insights from the ongoing EPSRC-funded ATRACT project, which is developing an AI enabled autonomous drone to support battlefield casualty triage by assisting frontline medics in the critical post trauma minutes that shape survival outcomes. Drawing on the 2026 article 'An AI ethics framework for a trustworthy autonomous drone system to support battlefield casualty triage' by Peter Lee, Tasweer Ahmad, Syed Waheed and Andrew Kenning, in AI and Ethics Vol. 6, the paper addresses a key underexplored question: how can abstract AI ethics principles be operationalised for a high stakes, safety critical socio technical system operated under high risk and extreme time pressure? The presentation will present a bespoke, project embedded ethics framework which translates high level AI ethics principles – human centricity, responsibility, explicability, bias and harm mitigation, and reliability – into concrete design, governance, and assurance practices. Central to the argument is that trustworthiness is not a property of algorithms alone, but emerges from documentation, oversight, auditability, and sustained human–AI teaming. The triage drone is deliberately designed as a decision support capability under meaningful human control, with responsibility for clinical judgement retained by qualified medics. The paper contributes practical guidance for researchers, engineers, and policymakers seeking to embed ethics into real world AI systems in defence and other safety critical domains.

Teaching Early-Stage Digital Design Reasoning Through an AI Design Workshop

Track: Emerging Research

Authors: Ross Brisco (University of Strathclyde), Kieran Gunn (University of Strathclyde)

Presenter: Ross Brisco (University of Strathclyde)

Abstract: This project presents a participatory AI design workshop developed to teach early-stage digital design reasoning through reflective and responsible AI use. Centred on a Product Design Assistant Tool, the workshop guided Engineering students in prompt engineering, iterative refinement, and critical evaluation of AI-generated outputs during conceptual product design activities. Students explored how changes in prompts, creativity settings, and contextual framing influenced design outcomes, encouraging deeper engagement with problem framing, constraints, and design justification. Findings suggest that workshop-based AI interventions can support AI literacy, reflective practice, and structured design thinking, positioning AI as a scaffold for learning rather than a replacement for design reasoning. We propose this as a workshop, demo or a poster to communicate the capabilities of the tool, lessons learned, and how we designed the workshop with students to support their learning.

HeartBERT: A self-supervised ECG embedding model for efficient and effective medical signal analysis

Track: World Class Research Track

Authors: Saeed Farzi (Southampton university)

Presenter: Saeed Farzi (Southampton university)

Abstract: This paper introduces HeartBERT, a self-supervised foundation model for ECG signal analysis. Inspired by RoBERTa and advances in natural language processing, HeartBERT reformulates ECG signals as synthetic language by applying signal quantisation, discretisation, and tokenisation. This enables transformer-based pre-training on large volumes of unlabelled ECG data, allowing the model to learn reusable contextual representations of physiological signal patterns without relying on costly manual annotation. The proposed model is designed as a general-purpose ECG representation backbone that can be fine-tuned for different downstream clinical tasks using limited labelled data. Its effectiveness is demonstrated through two clinically relevant applications: sleep-stage classification and heartbeat classification. Across these tasks, HeartBERT shows strong performance compared with baseline models, highlighting the value of domain-specific foundation modelling for biomedical time-series analysis. The paper further contrasts HeartBERT with GPT-4o, showing that general-purpose language models remain limited for specialised ECG interpretation without task-specific adaptation. Overall, this work demonstrates how foundation-model principles can be transferred from natural language to physiological signals, establishing HeartBERT as a reusable ECG modelling framework and opening a pathway toward broader biosignal foundation models, including future extensions to EEG and multimodal physiological monitoring.

Causal-Adapter: Taming Text-to-Image Diffusion for Faithful Counterfactual Generation

Track: World Class Research Track

Authors: Zhihua Liu (University of Edinburgh)

Presenter: Zhihua Liu (University of Edinburgh)

Abstract: Counterfactual image generation asks a simple but important question: what would an image look like if one meaningful factor were changed? For example, what would a brain scan look like if the ventricle were larger, or what would a face look like if the person were older? Current image editing models can often make visually realistic changes, but they may also alter unrelated details or follow superficial correlations rather than true cause-and-effect relationships. We introduce Causal-Adapter, a lightweight module that helps a frozen text-to-image model perform more faithful counterfactual editing. Instead of retraining the whole image generation model, our method injects causal information about attributes, such as age, gender, beard, or brain volume, into the model through a small adapter. This allows the model to change the intended attribute while preserving the identity and other important image details. Experiments on synthetic scenes, human faces, and medical brain images show that Causal-Adapter produces more accurate, realistic, and causally consistent edits than existing methods. This can make image generation more useful for scientific analysis, medical imaging research, and trustworthy visual reasoning.

ConceptSMILE: A Perturbation-Based Auditing Framework for Reliable Concept-Level Explanations in Retinal AI

Track: Emerging Research

Authors: Mohadeseh Mollapour (University of Hull)

Presenter: Mohadeseh Mollapour (University of Hull)

Abstract: This paper introduces ConceptSMILE, a groundbreaking framework for rigorously auditing the reliability of concept-level explanations in retinal AI. While concept-based explainable AI methods strive to enhance interpretability, the presence of clinically meaningful concepts does not guarantee reliable or trustworthy explanations. ConceptSMILE advances the original SMILE methodology by integrating controlled superpixel perturbations and locality-aware surrogate modeling to evaluate explanation reliability robustly. This innovative framework assesses how crucial retinal concepts such as lesions, blood vessels, and the optic disc respond to perturbations, measuring explanation reliability through key metrics including accuracy, fidelity, faithfulness, stability, and consistency. Preliminary experiments utilizing segmentation-based and vision-language explanation pathways reveal that ConceptSMILE can effectively pinpoint variations in explanation reliability across diverse model pathways, paving the way for more transparent, auditable, and clinically relevant retinal AI systems.

Whose Revolution? Rethinking AI, Belonging, and Sustainability in UK Higher Education

Track: Emerging Research

Authors: Tom Ritchie (University of Warwick)

Presenter: Tom Ritchie (University of Warwick)

Abstract: This provocation argues that AI equity, belonging in STEM, and sustainability are not three separate research problems. They are one. The UK research agenda funds them apart, convenes them separately, and produces partial solutions as a result. Drawing on a programme of work that began with a single interdisciplinary course and has since generated a worldwide teaching community, a national centre, an international quality assurance working group, and public scholarship reaching 16,000 or more listeners across 110 countries, this paper makes the case for a unified research agenda. It asks what the UK AI research community is missing by treating these conversations as distinct, and calls for the infrastructure and funding to bring them together.

Causal Ordering for Structure Learning from Time Series

Track: World Class Research Track

Authors: Pedro Sanchez (University of Edinburgh), Damian Machlanski (University of Edinburgh), Steven McDonagh (University of Edinburgh), Sotirios Tsaftaris (University of Edinburgh)

Presenter: Damian Machlanski (University of Edinburgh)

Abstract: Predicting causal structure from time series data is crucial for understanding complex phenomena in physiology, brain connectivity, climate dynamics, and socio-economic behaviour. Causal discovery in time series is hindered by the combinatorial complexity of identifying true causal relationships, especially as the number of variables and time points grows. A common approach to simplify the task is the so-called ordering-based methods. Traditional ordering methods inherently limit the representational capacity of the resulting model. In this work, we fix this issue by leveraging multiple valid causal orderings, instead of a single one as standard practice. We propose DOTS (Diffusion Ordered Temporal Structure), using diffusion-based causal discovery for temporal data. By integrating multiple orderings, DOTS effectively recovers the transitive closure of the underlying directed acyclic graph (DAG), mitigating spurious artifacts inherent in single-ordering approaches. We formalise the problem under standard assumptions such as stationarity and the additive noise model, and leverage score matching with diffusion processes to enable efficient Hessian estimation. Extensive experiments validate the approach. Empirical evaluations on synthetic and real-world datasets demonstrate that DOTS outperforms state-of-the-art baselines, offering a scalable and robust approach to temporal causal discovery. On synthetic benchmarks spanning d=3−6 variables, T=200−5,000 samples and up to three lags, DOTS improves mean window-graph F1 from 0.63 (best baseline) to 0.81. On the CausalTime real-world benchmark (Medical, AQI, Traffic; d=20−36), while baselines remain the best on individual datasets, DOTS attains the highest average summary-graph F1 while halving runtime relative to graph-optimisation methods. These results establish DOTS as a scalable and accurate solution for temporal causal discovery. Code is available at https://github.com/CHAI-UK/DOTS.

Autonomous agentic AI with policy adaptation for physics-informed spectral learning in Structural Health Monitoring

Track: World Class Research Track

Authors: Anshu Sharma (University of Strathclyde), Basuraj Bhowmik (University of Strathclyde)

Presenter: Basuraj Bhowmik (University of Strathclyde)

Abstract: This study develops a physics-grounded agentic artificial intelligence (AI) framework for Structural Health Monitoring (SHM) that integrates perception, cognition, action, and reflection with structural dynamics and stochastic system identification. The framework differs from conventional machine learning or deep learning pipelines by optimizing long-horizon monitoring objectives, autonomously adapting sensing and inference to meet safety-critical decision requirements. The methodology establishes a direct bridge from multi-degree-offreedom structural dynamics to output-only spectral identification through Frequency Domain Decomposition (FDD), and subsequently to Bayesian inference of damage hypotheses and adaptive policy optimization. The proposed framework introduces three main contributions. First, it represents the explicit integration of agentic AI with SHM, embedding reasoning, planning, and learning within a closed-loop structure that is physically interpretable. Second, it provides a rigorous connection between computer science formulations such as partially observable Markov decision processes and reinforcement learning with the established domain of linear structural dynamics and modal analysis. Third, it demonstrates reproducibility and effectiveness through a numerical study on a five-degree-of-freedom shear building subjected to Gaussian white noise excitation. The results indicate high-fidelity modal identification and accurate localization of stiffness loss, with reliability assessed against theoretical baseline modes using the Modal Assurance Criterion (MAC). Further, the robustness of the agentic AI is validated using the ASCE SHM benchmark structure. The findings point toward a new generation of monitoring systems that are physics-informed, uncertainty-aware, and capable of resilient and adaptive operation across interconnected infrastructure networks.

Achieving collective welfare in multi-agent reinforcement learning via suggestion sharing

Track: World Class Research Track

Authors: Yue Jin (University of Northampton), Shuangqing Wei (Louisiana State University), Giovanni Montana (University of Warwick)

Presenter: Yue Jin (University of Northampton)

Abstract: In human society, the conflict between self-interest and collective well-being often obstructs efforts to achieve shared welfare. Related concepts like the Tragedy of the Commons and Social Dilemmas frequently manifest in our daily lives. As artificial agents increasingly serve as autonomous proxies for humans, we propose a novel multi-agent reinforcement learning (MARL) method to address this issue—learning policies to maximise collective returns even when individual agents' interests conflict with the collective one. Unlike traditional cooperative MARL solutions that involve sharing rewards, values, and policies or designing intrinsic rewards to encourage agents to learn collectively optimal policies, we propose a novel MARL approach where agents exchange action suggestions. Our method reveals less private information compared to sharing rewards, values, or policies, while enabling effective cooperation without the need to design intrinsic rewards. Our algorithm is supported by our theoretical analysis that establishes a bound on the discrepancy between collective and individual objectives, demonstrating how sharing suggestions can align agents' behaviours with the collective objective. Experimental results demonstrate that our algorithm performs competitively with baselines that rely on value or policy sharing or intrinsic rewards.

Partial Action Replacement: Tackling Distribution Shift in Offline MARL

Track: World Class Research Track

Authors: Yue Jin (University of Northampton), Giovanni Montana (University of Warwick)

Presenter: Yue Jin (University of Northampton)

Abstract: Offline multi-agent reinforcement learning (MARL) is severely hampered by the challenge of evaluating out-of-distribution (OOD) joint actions. Our core finding is that when the behavior policy is factorized—a common scenario where agents act fully or partially independently during data collection—a strategy of partial action replacement (PAR) can significantly mitigate this challenge. PAR updates a single or part of agents' actions while the others remain fixed to the behavioral data, reducing distribution shift compared to full joint-action updates. Based on this insight, we develop Soft-Partial Conservative Q-Learning (SPaCQL), using PAR to mitigate OOD issue and dynamically weighting different PAR strategies based on the uncertainty of value estimation. We provide a rigorous theoretical foundation for this approach, proving that under factorized behavior policies, the induced distribution shift scales linearly with the number of deviating agents rather than exponentially with the joint-action space. This yields a provably tighter value error bound for this important class of offline MARL problems. Our theoretical results also indicate that SPaCQL adaptively addresses distribution shift using uncertainty-informed weights. Our empirical results demonstrate SPaCQL enables more effective policy learning, and manifest its remarkable superiority over baseline algorithms when the offline dataset exhibits the independence structure.

Beyond Second-Order: Cubic-Quartic Regularization Methods for Nonconvex Optimization in Machine Learning

Track: World Class Research Track

Authors: Kate Wenqi Zhu (Oxford), Coralia Cartis (Oxford)

Presenter: Kate Wenqi Zhu (Oxford)

Abstract: Nonconvex optimization is fundamental to modern machine learning — from training deep neural networks and generative models to reinforcement learning and neural architecture search. While first-order methods (SGD, Adam) and second-order methods (natural gradient, K-FAC) remain dominant in practice, they face well-known limitations: slow convergence near saddle points, sensitivity to ill-conditioning, and limited exploitation of loss landscape geometry. High-order tensor methods that incorporate third-order (and beyond) derivative information offer a principled path forward, achieving provably faster local convergence and optimal worst-case evaluation complexity, while capturing richer geometric structure of the loss landscape. A central barrier to deploying these methods has been the computational cost of their core subproblem: minimizing a nonconvex multivariate polynomial of degree three or higher, regularised by an appropriate power of the step size. In this work, we propose the CQR (Cubic-Quartic Regularization) framework, which addresses this bottleneck for third-order tensor methods by replacing the expensive subproblem with a sequence of tractable local models that incorporate both cubic and quartic terms. The cubic term approximates local tensor information while the quartic term provides regularisation and ensures well-posedness. We derive necessary and sufficient optimality conditions that fully characterise the global minimisers of these models, reduce them to secular equations solvable via nonlinear eigenvalue techniques, and prove that a CQR variant achieves the optimal evaluation complexity of $O(\epsilon^{-3}/^{2})$. Numerical experiments demonstrate that CQR is competitive with standard subproblem solvers on typical instances and superior on ill-conditioned problems with special structure — a regime frequently encountered in deep learning, where loss Hessians exhibit extreme spectral spread. Our framework lays the groundwork for making high-order optimisation methods practical for AI applications, particularly in settings where the geometry of the loss surface matters most.

Self-Guided Integrated Gradient Method for Attribution

Track: World Class Research Track

Authors: Sabrina Henry (Heriot-Watt University), Alice Ruget (Heriot-Watt University), Stirling Scholes (University of Strathclyde), Jonathan Leach (Heriot-Watt University)

Abstract: Explaining the decisions of deep neural networks is essential for building trust in AI systems, particularly when deployed in sensitive domains such as healthcare, security, and autonomous transport. At the same time, the growing complexity of these models makes their internal reasoning difficult to understand. In response, research in eXplainable Artificial Intelligence (XAI) has emerged to develop methods that help humans interpret, and when appropriate, trust, the decisions of modern AI systems. Path-based attribution methods such as Integrated Gradients estimate how each input feature contributes to a model's prediction by integrating model gradients along a path from a baseline to the input image. However, such methods require user-defined baselines and often accumulate noisy gradients from saturated regions of the prediction landscape. To address these limitations, we propose the Self-guided Integrated Gradient Method for Attribution (SIGMA), a baseline-free path-based method that stochastically explores the model's confidence landscape to identify input features responsible for the collapse of class confidence. By following the model's decision boundary, SIGMA produces interpretable and reliable attributions without requiring reference inputs or access to a model's internal representation layers. Evaluations across diverse architectures, including Vision Transformers, and computer vision datasets in healthcare and security, demonstrate that SIGMA provides spatially coherent attribution maps with strong faithfulness to the model's internal reasoning.

Higher-Order Feature Attribution: Bridging Statistics, Explainable AI, and Topological Signal Processing

Track: World Class Research Track

Authors: Kurt Butler (The University of Edinburgh), Guanchao Feng (Stony Brook University), Petar Djuric (Stony Brook University)

Presenter: Kurt Butler (The University of Edinburgh)

Abstract: Feature attributions are post-training analysis methods that assess how various input features of a machine learning model contribute to an output prediction. Their interpretation is straightforward when features act independently, but it becomes less clear when the predictive model involves interactions, such as multiplicative relationships or joint feature contributions. In this work, we propose a general theory of higher-order feature attribution, which we develop on the foundation of Integrated Gradients (IG). This work extends existing frameworks in the literature on explainable AI. When using IG as the method of feature attribution, we discover natural connections to statistics and topological signal processing. We provide several theoretical results that establish the theory, and we validate our theory on a few examples

From Lab to Wrist: A Responsible AI Framework for Wearable Ground Reaction Force Estimation

Track: Emerging Research

Authors: Parvin Ghaffarzadeh (University of Hull), Debarati Chakraborty (University of Hull), Koorosh Aslansefat (University of Hull), Ali Dostan (Nottingham Trent University), Yiannis Papadopoulos (University of Hull)

Abstract: Wearable AI systems for ground reaction force (GRF) estimation could support bone health monitoring, rehabilitation, and athletic load management, yet a systematic evidence map of 98 primary studies shows the field is structurally misaligned with deployment: subject-independent validation appears in only 6 % of studies; consumer smartwatch evaluation and time-resolved explainabil- ity were rarely reported, each appearing in only one study un- der our review criteria. We contribute the Translation Readiness Framework (TRF), a five-dimension deployment-readiness diagnos- tic grounded in responsible AI principles, and apply it to identify evaluation, infrastructure, and governance gaps that must close before wearable GRF systems reach health contexts. A proof-of- concept system, GRFNet-MultiScale, demonstrates the feasibility of selected translation-readiness criteria on consumer hardware (LOSO r =0.798, dual Apple Watch, N =10), with on-device iPhone inference, TimeSHAP temporal attributions, and an explicit pre- defined scope boundary distinguishing validated cyclic locomotion from out-of-scope impact tasks and clinical inference. We addition- ally outline a concept-level responsible-AI extension for runtime scope monitoring, intended to flag out-of-context use before descriptive GRF estimates are interpreted.

A Survey of Expectation-Maximization Methods for Policy Optimization in Deep Reinforcement Learning

Track: Emerging Research

Authors: Adrian Mircea Nenu (The University of Manchester), Satya Prakash Dash (The University of Manchester), Yingxiao Huo (The University of Manchester), Yihe Zhou (The University of Manchester), Brendan Bennett (The University of Manchester), Beining Zhang (The University of Manchester), Victor Yeom Song (Aalto University), Samuel Kaski (The University of Manchester), Mingfei Sun (The University of Manchester)

Presenter: Adrian Mircea Nenu (The University of Manchester)

Abstract: Expectation-Maximization (EM) underlies a growing set of policyoptimization methods across continuous control and robotics, offline reinforcement learning, preference-based language-model alignment, and self-training for reasoning with verifiable rewards. This survey unifies these methods under EM-based policy optimization, a single materialize-then-project loop. The E-step forms a non-parametric target by maximizing a KL-regularized advantage, tilting a reference policy by an exponentiated advantage to form a Boltzmann target, then the M-step projects that target back onto the parametric policy. We group the methods into a weighted-MLE loop with roots in robotics and continuous control and its modifications centered around exact-projection variants, non-weighted- MLE M-steps, and zero-temperature self-training. We then catalog the code-level optimizations based on the implementations and empirical results reported by surveyed methods.

GenGPLight: A Generalized Genetic Programming Framework for Traffic Signal Control in Arbitrary Intersection Structures

Track: World Class Research Track

Authors: Sahar Kianian (University of Exeter), Edward Keedwell (University of Exeter)

Abstract: This paper introduces GenGPLight, a generalised framework for optimizing traffic signal control across intersections with arbitrary structures, sizes, and signal configurations. GenGPLight is a learning-based optimisation approach that uses genetic programming to automatically synthesise urgency functions for adaptive signal timing, which adjusts timings based on queue lengths detected by sensors or cameras. It leverages a flexible phase-level feature representation derived from aggregated lane-level traffic metrics, enabling the evolution of interpretable urgency-based control strategies. The framework is evaluated on real-world traffic networks, including both structured and unstructured scenarios, and benchmarked against established baselines. Results show that GenGPLight achieves robust performance, improved traffic efficiency, and strong generalisation, highlighting its suitability for deployment in complex urban environments.

MCCE: A Framework for Multi-LLM Collaborative Search in Discrete Spaces with Similarity-Filtered Preference Learning

Track: World Class Research Track

Authors: Nian Ran (University of Manchester)

Presenter: Nian Ran (University of Manchester)

Abstract: Multi-objective discrete optimization problems, such as molecular design, pose significant challenges due to their vast and unstructured combinatorial spaces. Traditional evolutionary algorithms often get trapped in local optima, while expert knowledge can provide crucial guidance for accelerating convergence. Large language models (LLMs) offer powerful priors and reasoning ability, making them natural optimizers when expert knowledge matters. However, closed-source LLMs, though strong in exploration, cannot update their parameters and thus cannot internalize experience. Conversely, smaller open models can be continually fine-tuned but lack broad knowledge and reasoning strength. We introduce Multi-LLM Collaborative Co-evolution (MCCE), a hybrid framework that unites a frozen closed-source LLM with a lightweight trainable model. The system maintains a trajectory memory of past search processes; the small model is progressively refined via reinforcement learning, with the two models jointly supporting and complementing each other in global exploration. Unlike model distillation, this process enhances the capabilities of both models through mutual inspiration. Experiments on multi-objective drug design benchmarks show that MCCE achieves state-of-the-art Pareto front quality and consistently outperforms baselines. These results highlight a new paradigm for enabling continual evolution in hybrid LLM systems, combining knowledge-driven exploration with experience-driven learning. The code of MCCE is available on https://github.com/lzz-z/MCCE

Structural imaging predictors of ketamine response in treatment-resistant depression: a machine learning approach

Track: World Class Research Track

Authors: Linda Bryant (King's College London)

Presenter: Linda Bryant (King's College London)

Abstract: Treatment-resistant depression (TRD) remains a major clinical challenge, with ketamine emerging as a rapid-acting intervention despite substantial variability in individual treatment response. Current clinical decision-making largely relies on trial-and-error prescribing, highlighting the need for predictive tools that can support personalized treatment selection. In this study, we developed and externally validated a machine learning model using pre-treatment structural MRI data to predict antidepressant response to ketamine in patients with TRD. Regional gray matter volume measures derived from structural neuroimaging were used to train a support vector classifier within a repeated nested cross-validation framework. Model generalizability was evaluated in independent ketamine-treated cohorts and specificity was assessed using a saline-treated control cohort. The model achieved a balanced accuracy of 72.2% in the discovery sample and retained significant predictive performance during external validation. Frontal cortical regions were associated with treatment response, while cerebellar regions predicted non-response. Importantly, predictive performance dropped to chance levels in the saline-treated cohort, supporting pharmacological specificity. This work demonstrates how artificial intelligence can leverage neuroimaging data to generate clinically meaningful predictions in psychiatry. The findings provide evidence that structural brain biomarkers may help guide individualized treatment selection while offering insight into the neurobiological mechanisms underlying ketamine response. More broadly, the study highlights the potential of machine learning approaches to advance precision medicine in mental health.

Digital manufacturing platform for syngas production from waste wood

Track: Emerging Research

Authors: Waqar Muhammad Ashraf (University College London), Massimiliano Materazzi (University College London), Vivek Dua (University College London)

Presenter: Waqar Muhammad Ashraf (University College London)

Abstract: The heterogeneous nature of waste biomass feedstock and fluctuating process conditions are key barriers in maintaining stability and quality of syngas production from waste feedstock. Motivated by these challenges, this paper develops a machine learning (ML)-integrated optimisation framework which is implemented for maximising net energy gain from the syngas produced from plasma-assisted gasification of waste wood. Optimisation results show that under a 150 kW plasma power consumption constraint, H2 (37.2%) and CO (39.4%) can be produced for wood feed rate of 30.3 kg/h and steam feed rate of 26.7 kg/h. The developed ML-powered optimisation framework is integrated into a graphical-user interface platform, offering a simulation environment assisted real-time process control for waste to syngas production for circular economy initiatives.

Profit-Driven Artificial Intelligence, Digital Health, and the Risk of Losing Public Trust

Track: Emerging Research

Authors: MARJORY DA COSTA ABREU (Sheffield Hallam University)

Abstract: In this paper, we will discuss how Profit-driven artificial intelligence (Profit-AI) is increasingly shaping the future of digital health innovation within the NHS and wider UK healthcare system. While initiatives such as the EPSRC South Yorkshire Digital Health Hub are demonstrating the potential for ethical and inclusive innovation, significant risks remain when Digital Health Innovation is on the table and the use of ProfitAI is considered. Thus, we will examine three critical challenges associated with using ProfitAI in digital health: the widening of health inequalities through biased systems and digital exclusion; the growing dependence of the NHS on private technological infrastructure, risking loss of national sovereignty; and the erosion of public trust through weak governance and limited transparency. We argue that ethical, publicly accountable ProfitAI governance is essential for sustainable healthcare innovation.

AI-Discovered Physical Identities for Secure Authentication Using Optical Material Signatures

Track: Emerging Research

Authors: Syeda Ramsha Ali (University of Soutampton), Glen Mutinda (University of Soutampton), Nema Abdelazim (University of Soutampton)

Presenter: Syeda Ramsha Ali (University of Soutampton)

Abstract: Artificial intelligence (AI) has achieved remarkable success in tasks such as image recognition, language understanding, content generation and predictive modelling. Despite these advances, most AI systems remain fundamentally designed to learn categories, labels and statistical relationships within data. They are highly effective at answering questions such as What is this? but are rarely designed to answer a more fundamental question: What makes this specific physical object unique itself? This limitation becomes particularly significant in authentication and anti-counterfeiting applications. Modern supply chains increasingly require mechanisms to verify the authenticity of products, components, pharmaceuticals, electronic devices and critical infrastructure. Existing authentication approaches often depend on predefined identifiers, handcrafted features and supervised machine learning (ML) models trained using both genuine and counterfeit samples. However, in real-world counterfeit examples are frequently unavailable, continuously evolving, or intentionally designed to mimic genuine products. As a result, conventional AI systems struggle to adapt to emerging threats. This project explores a new research direction termed Physical Identity Discovery, which investigates whether AI can autonomously learn the intrinsic identity of a physical object directly from scientific measurements. Rather than classifying an object into a predefined category, the objective is to discover stable identity representations that remain consistent across repeated measurements while preserving the uniqueness that distinguishes one object from another. To investigate this concept, quantum dot (QD) optical tags are used as a model platform. These materials generate rich excitation-emission matrix (EEM) signatures containing multidimensional optical information arising from microscopic variations in material composition and structure. Such measurements provide an ideal testbed for studying how AI can extract identity information from complex physical systems. The central hypothesis is that representation learning can transform high-dimensional scientific measurements into compact latent identity spaces that capture intrinsic object uniqueness. If successful, authentication could be achieved by verifying

consistency with a learned identity representation rather than relying on explicit counterfeit examples. This would represent a shift from classification-based authentication towards identity-based authentication. Beyond optical security tags, the implications of this research extend to broader scientific and engineering domains. The ability to learn physical identity from measurement data could support trustworthy AI systems for advanced manufacturing, secure supply chains and future digital trust infrastructures. Ultimately, this work seeks to establish the foundations of a new AI capability in which ML systems move beyond recognising categories and begin learning the unique physical identities of objects from experimental data. Such capabilities may contribute to the next generation of trustworthy AI systems capable of reasoning about authenticity, uniqueness and bases for complex real-world environments.

Clear, Compelling Arguments: Rethinking the Foundations of Frontier AI Safety Cases

Track: World Class Research Track

Authors: Shaun Feakins (University of York), Ibrahim Habli (University of York), Phillip Morgan (University of York)

Presenter: Shaun Feakins (University of York)

Abstract: This paper was accepted at IASEAI'26, due to be published in AAAI proceedings, and a follow-up was presented as a poster at the CHAI Workshop, UC Berkeley. Abstract: This paper contributes to the nascent debate around safety cases for frontier AI systems. Safety cases are structured, defensible arguments that a system is acceptably safe to deploy in a given context. Historically, they have been used in safety-critical industries, such as aerospace, nuclear or automotive. As a result, safety cases for frontier AI have risen in prominence, both in the safety policies of leading frontier developers and in international research agendas proposed by leaders in generative AI, such as the Singapore Consensus on Global AI Safety Research Priorities and the International AI Safety Report. This paper appraises this work. We note that research conducted within the alignment community which draws explicitly on lessons from the assurance community has significant limitations. We therefore aim to rethink existing approaches to alignment safety cases. We offer lessons from existing methodologies within safety assurance and outline the limitations involved in the alignment community's current approach. Building on this foundation, we present a case study for a safety case focused on Deceptive Alignment and CBRN capabilities, drawing on existing, theoretical safety case "sketches" created by the alignment safety case community. Overall, we contribute holistic insights from the field of safety assurance via rigorous theory and methodologies that have been applied in safety-critical contexts. We do so in order to create a better foundational framework for robust, defensible and useful safety case methodologies which can help to assure the safety of frontier AI systems.

Predicting future Dementia diagnosis from routine clinical brain MRI: A Scotland-wide pilot

Track: Emerging Research

Authors: Dorian Gouzou (University of Edinburgh), Michael P. J. Camilleri (University of Edinburgh), Muthu R. K. Mookiah (University of Dundee), Maria Valdes Hernandez (University of Edinburgh), Sotirios Tsaftaris (University of Edinburgh), Douglas Steele (University of Dundee), Alexander Doney (University of Dundee), Emanuele Trucco (University of Dundee), William Whiteley (University of Edinburgh)

Presenter: Michael P. J. Camilleri (University of Edinburgh)

Abstract: Dementia diagnosis is often delayed, motivating methods that identify risk earlier using routine clinical data. We investigated whether structural brain MRI acquired during routine care can predict subsequent all-cause dementia diagnosis using a Scotland-wide imaging cohort linked to health records. We compared Logistic Regression, Support Vector Machine and Deep Learning predictors on MRI scans from 15.5K patients. On a held-out test set, LR achieved the highest AUROC of 0.70 and balanced accuracy of 0.64, outperforming the other models despite using lower-dimensional anatomical features. These findings suggest that routine imaging contains signal associated with later dementia diagnosis, although future work is needed to address censoring and calibration.

Overview of the CLPsych 2026 Shared Task: Capturing and Characterizing Mental Health Changes through Social Media Timeline Dynamics

Track: World Class Research Track

Authors: Iqra Ali (Queen Mary University of London), Talia Tseriotou (Queen Mary University of London), Guy Dvir (Bar Ilan University), Callum Chan (University of Ottawa), Yuxiang Zhou (Queen Mary University of London), Juan Antonio Lossio-Ventura (National Institutes of Health), Ayal Klein (Bar Ilan University), Aya Shamir (Bar Ilan University), Dan Sayda (Bar Ilan University), Aya Zirikly (The George Washington University, Johns Hopkins University), Diana Inkpen (University of Ottawa), Dana Atzil-Slonim (Bar Ilan University), Maria Liakata (Queen Mary University of London, The Alan Turing Institute)

Presenter: Iqra Ali (Queen Mary University of London)

Abstract: We provide an overview of the CLPsych 2026 Shared Task, which focuses on capturing and characterizing mental health dynamics from social media timelines through structured modeling of self-states. This year advances the longitudinal paradigm set by prior CLPsych shared tasks (2022, 2025), by integrating fine-grained psychological representation using the MIND framework. The task is organized into three main components: (1) post-level identification of adaptive and maladaptive self-states through ABCD elements and sub-elements, along with estimation of their presence; (2) timeline-level detection of Moments of Change, including both abrupt switches and gradual escalations based on ABCd element and sub-element combinations; and (3) sequence-level modeling, involving summarization of change processes over time and identification of recurrent dynamic signatures.

Low Rank properties and architectures in small Image Classification Neural Networks

Track: World Class Research Track

Authors: Mark Sandler (Queen Maru University of London)

Abstract: This paper explores training of small, LeNet Deep Learning models with an emphasis on the rank of weight matrices. The first part of the paper examines conventional models, using the Singular Value Decomposition (SVD) to decompose the LeNet linear layers when trained with simple, standard image data sets, primarily, MNIST. The second part of the paper examines modified models in which the 2 linear layers are replaced with a specific structure that supports and emphasises low rank implementation: these are named Sum of Rank One (SoRO) layers. It is found particularly useful to measure the angular evolution of the left and right singular vector matrices (and their equivalent in SoRO layers). We find some surprising behaviour in conventional LeNets when trained beyond the point of negligible training loss. We further find that the linear layers in SoRO-edited LeNets can be rank-reduced by more than 90% while maintaining performance. Additionally, we see substantially different behaviour in the evolution of singular vectors in SoRO layers, during training, which brings insight to how deep learning works. This paper has been accepted at Eusipco 2026, Europe's premier signal processing conference

A Universal Nearest-Neighbor Estimator for Intrinsic Dimensionality

Track: Emerging Research

Authors: Omer Bobrowski (Queen Mary University of London), Primoz Skraba (Queen Mary University of London), Gesine Reinert (University of Oxford), Eng-Jon Ong (Queen Mary University of London)

Presenter: Eng-Jon Ong (Queen Mary University of London)

Abstract: Estimating the intrinsic dimensionality (ID) of data is a fundamental problem in AI, machine learning and data analysis. The ID represents the true number of degrees of freedom our high-dimensional observations. Existing approaches often rely on specific distributional models, or local approximations that can degrade when their assumptions are violated. We introduce L2N2, a simple nearest-neighbor-based estimator of intrinsic dimensionality. The estimator is built from ratios of nearest-neighbor distances and requires only the computation of sample averages. Theoretically, under broad conditions, the estimator converges to the true intrinsic dimension independently of the sampling distribution. Experimental results on synthetic manifolds and real-world datasets demonstrate that L2N2 achieves state-of-the-art performance while remaining computationally efficient.

Bayesian Sound Source Localisation of Free Flying Bats with a Gaussian Process Trajectory Prior

Track: Emerging Research

Authors: Stephen Ellis (University of Oxford)

Presenter: Stephen Ellis (University of Oxford)

Abstract: Locating and tracking animals from their vocalisations has proven itself indispensable for ecological research; with applications ranging from unbiased population estimation to kinematic and acoustic quantification. Near-field sound source localisation techniques infer the location of vocalising animals by estimating the signal propagation delay across a spatially distributed array of acoustic receivers. In this work we tackle the growing need for robust uncertainty quantification by considering a Bayesian approach to sound source localisation and trajectory estimation of the common pipistrelle bat. We estimate the posterior distribution on sound source locations via a general-purpose Markov chain Monte Carlo (MCMC) sampler, extending this model to infer continuous flight trajectories using a smooth Gaussian process prior. The resulting tool effectively captures measurement uncertainty, producing a principled probabilistic estimate of sound source localisation. We demonstrate the efficacy of this approach for field deployments using an array of five GPS synchronised AudioMoth receivers to track free flying bats, achieving 25cm accuracy.

Towards Transparent AI Agents for Air Traffic Control

Track: World Class Research Track

Authors: Elhassan Mohamed (The Alan Turing Institute), Ben Carvell (NATS), Rob Procter (University of Warwick & The Alan Turing Institute), Eseoghene Benjamin (The Alan Turing Institute), George De Ath (University of Exeter), Richard Everson (University of Exeter & The Alan Turing Institute)

Presenter: Elhassan Mohamed (The Alan Turing Institute)

Abstract: Advances in the use of artificial intelligence agents for air traffic control (ATC) have the potential to reshape modern aviation operations. Despite these advances, the adoption of such agents in safety-critical ATC applications remains limited and this is, in part, due to their lack of transparency. The transparency of such agents is crucial for humans to understand why a specific action is advised or to reason about the agent's behaviour and assess their trustworthiness in real-time. The simpler the agent, the more transparent it can be, but this comes at the cost of flexibility in agent performance. We propose, illustrate, and critically examine mechanisms for providing AI agents for ATC with interpretability and explainability. We focus on agents ranging from simple rule-based systems to optimisation-based and reinforcement-learning agents, and we report initial qualitative feedback from operational ATCOs on prototype explainability mechanisms.

Online Action-Stacking Improves Reinforcement Learning Performance for Air Traffic Control

Track: World Class Research Track

Authors: Ben Carvell (NATS), George De Ath (University of Exeter), Eseoghene Benjamin (The Alan Turing Institute), Richard Everson (University of Exeter)

Presenter: Eseoghene Benjamin (The Alan Turing Institute)

Abstract: We introduce online action-stacking, an inference-time wrapper for reinforcement learning policies that produces realistic air traffic control commands while allowing training on a much smaller discrete action space. Policies are trained with simple incremental heading or level adjustments, together with an action-damping penalty that reduces instruction frequency and leads agents to issue commands in short bursts. At inference, online action-stacking compiles these bursts of primitive actions into domain-appropriate compound clearances. Using Proximal Policy Optimisation and the BluebirdDT digital twin platform, we train agents to navigate aircraft along lateral routes, manage climb and descent to target flight levels, and perform two-aircraft collision avoidance under a minimum separation constraint. In our lateral navigation experiments, action stacking greatly reduces the number of issued instructions relative to a damped baseline and achieves comparable performance to a policy trained with a 37-dimensional action space, despite operating with only five actions. These results indicate that online action-stacking helps bridge a key gap between standard reinforcement learning formulations and operational ATC requirements, and provides a simple mechanism for scaling to more complex control scenarios.

Human-in-the-Loop Testing of AI Agents for Air Traffic Control with a Regulated Assessment Framework

Track: World Class Research Track

Authors: Ben Carvell (NATS), Marc Thomas (NATS), Andrew Pace (NATS), Christopher Dorney (NATS), George De Ath (University of Exeter), Richard Everson (University of Exeter), Nick Pepper (The Alan Turing Institute), Adam Keane (The Alan Turing Institute), Samuel Tomlinson (NATS), Richard Cannon (NATS)

Presenter: Ben Carvell (NATS)

Abstract: We present a rigorous, human-in-the-loop evaluation framework for assessing the performance of AI agents on the task of Air Traffic Control, grounded in a regulator-certified simulator-based curriculum used for training and testing real-world trainee controllers. By leveraging legally regulated assessments and involving expert human instructors in the evaluation process, our framework enables a more authentic and domain-accurate measurement of AI performance. This work addresses a critical gap in the existing literature: the frequent misalignment between academic representations of Air Traffic Control and the complexities of the actual operational environment. It also lays the foundations for effective future human-machine teaming paradigms by aligning machine performance with human assessment targets.

Artificial intelligence for cultural heritage research: the challenges in UK copyright law and policy

Track: World Class Research Track

Authors: Paula Westenberger (Brunel University of London), Despoina Farmaki (City St George's, University of London)

Presenter: Paula Westenberger (Brunel University of London)

Abstract: Artificial intelligence (AI) is revolutionising our relationship with cultural heritage, enhancing access to, engagement with and preservation of collections and heritage sites. AI is also being used as a valuable research tool in the context of heritage collections. However, as materials protected by copyright may be used in AI development, training and use, copyright law can become an obstacle to important AI deployments in the heritage sector, an area which is currently understudied from the United Kingdom (UK) perspective. This article explores the intricate interplay between cultural heritage, AI and copyright law, demonstrating the main copyright law and policy challenges facing cultural heritage professionals and researchers in using AI in the UK for heritage research. It highlights the complexity and uncertainties as regards the current Text and Data Mining exception in the UK Copyright, Designs and Patents Act 1988 (UK CDPA), emphasising the need for an improved legal framework that balances copyright protection with the benefits of AI for cultural heritage research and management. It also reveals the underrepresentation of the heritage sector in AI regulation and copyright policy discussions in the UK. This exploration underscores the imperative for an inclusive policy dialogue that considers the perspectives and evidence of the cultural heritage sector in its full breadth and diversity (including related researchers) in shaping copyright law reform and AI regulation, and for further research to be carried out in this field.

The Alignment Bottleneck in Decomposition-Based Claim Verification

Track: World Class Research Track

Authors: Mahmud Elahi Akhter (Queen Mary University of London), Federico Ruggeri (University of Bologna), Iman Munire Bilal (University of Warwick), Robert Procter (University of Warwick), Maria Liakata (Queen Mary University of London)

Presenter: Mahmud Elahi Akhter (Queen Mary University of London)

Abstract: Structured claim decomposition is often proposed as a solution for verifying complex, multi-faceted claims, yet empirical results have been inconsistent. We argue that these inconsistencies stem from two overlooked bottlenecks: evidence alignment and sub-claim error profiles. To better understand these factors, we introduce a new dataset of real-world complex claims, featuring temporally bounded evidence and human-annotated sub-claim evidence spans. We evaluate decomposition under two evidence alignment setups: Sub-claim Aligned Evidence (SAE) and Repeated Claim-level Evidence (SRE). Our results reveal that decomposition brings significant performance improvement only when evidence is granular and strictly aligned. By contrast, standard setups that rely on repeated claim-level evidence (SRE) fail to improve and often degrade performance as shown across different datasets and domains (PHEMEPlus, MMM-Fact, COVID-Fact). Furthermore, we demonstrate that in the presence of noisy sub-claim labels, the nature of the error ends up determining downstream robustness. We find that conservative "abstention" significantly reduces error propagation compared to aggressive but incorrect predictions. These findings suggest that future claim decomposition frameworks must prioritize precise evidence synthesis and calibrate the label bias of sub-claim verification models.

A Probabilistic Digital Twin of UK En Route Airspace for Training and Evaluating AI Agents for Air Traffic Control

Track: World Class Research Track

Authors: George De Ath (The University of Exeter), Adam Keane (The Alan Turing Institute), Ben Carvell (NATS), Richard Everson (The University of Exeter), Marc Thomas (NATS), Richard Cannon (NATS), Amy Hodgkin (The Alan Turing Institute), Dewi Gould (The Alan Turing Institute), Edward Henderson (The Alan Turing Institute), Lynge Lauritsen (The Alan Turing Institute), Christos Vlahos (The Alan Turing Institute), Alvaro Sierra-Castro (NATS), John Korna (NATS), Nick Pepper (The Alan Turing Institute)

Presenter: Marc Thomas (NATS)

Abstract: This paper presents the first probabilistic Digital Twin of operational en route airspace, developed for the London Area Control Centre. The Digital Twin is intended to support the development and rigorous human-in-the-loop evaluation of AI agents for Air Traffic Control (ATC), providing a virtual representation of real-world airspace that enables safe exploration of higher levels of ATC automation. This paper makes three significant contributions: firstly, we demonstrate how historical and live operational data may be combined with a probabilistic, physics-informed machine learning model of aircraft performance to reproduce real-world traffic scenarios, while accurately reflecting the level of uncertainty inherent in ATC. Secondly, we develop a structured assurance case, following the Trustworthy and Ethical Assurance framework, to provide quantitative evidence for the Digital Twin's accuracy and fidelity. This is crucial to building trust in this novel technology within this safety-critical domain. Thirdly, we describe how the Digital Twin forms a unified environment for agent testing and evaluation. This includes fast-time execution (up to x200 real-time), a standardised Python-based "gym" interface that supports a range of AI agent designs, and a suite of quantitative metrics for assessing performance. Crucially, the framework facilitates competency-based assessment of AI agents by qualified Air Traffic Control Officers through a Human Machine Interface. We also outline further applications and future extensions of the Digital Twin architecture.

Air Traffic Controller Task Demand via Graph Neural Networks: An Interpretable Approach to Airspace Complexity

Track: World Class Research Track

Authors: Edward Henderson (The Alan Turing Institute), Dewi Gould (The Alan Turing Institute), Richard Everson (The Alan Turing Institute), George De Ath (University of Exeter), Nick Pepper (The Alan Turing Institute)

Presenter: Edward Henderson (The Alan Turing Institute)

Abstract: Real-time assessment of near-term Air Traffic Control Officer (ATCO) task demand is a critical challenge in an increasingly crowded airspace, as existing complexity metrics often fail to capture nuanced operational drivers beyond simple aircraft counts. This work introduces an interpretable Graph Neural Network (GNN) framework to address this gap. Our attention-based model predicts the number of upcoming clearances, the instructions issued to aircraft by ATCOs, from interactions within static traffic scenarios. Crucially, we derive an interpretable, per-aircraft task demand score by systematically ablating aircraft and measuring the impact on the model's predictions. Our framework significantly outperforms an ATCO-inspired heuristic and is a more reliable estimator of scenario complexity than established baselines. The resulting tool can attribute task demand to specific aircraft, offering a new way to analyse and understand the drivers of complexity for applications in controller training and airspace redesign.

Perspectivism and purposiveness toward an ancestral future: cutting AI's guts open

Track: World Class Research Track

Authors: Beatrice Bonami (University of Edinburgh)

Presenter: Beatrice Bonami (University of Edinburgh)

Abstract: Artificial Intelligence (AI) narratives are commonly set on risk – which, this paper argues, leads to a recursive loop mirroring the narratives found in science fiction where AI threatens human existence. These tales are often shaped by the fear that AI will eventually harm or rival human authenticity and organic situatedness (Haraway 1988). Such a loop contributes to the formation of black boxes (Latour 1999) that, despite ensuring data security, secrecy, and privacy, serve the interests of the private sector. To challenge these black boxes, I draw on the ontological turn in anthropology (Bessire and Bond 2014; Heywood 2017) and black feminist decoloniality (Ricaurte 2019). These perspectives offer a space to non-Western Southern epistemologies through which AI can be understood beyond the constraints of risk construction. Two key concepts emerge in this context – perspectivism (Castro 2015; Cordeiro 2022) and purposiveness (Dladla 2017) – which serve as analytical models for exposing the inner workings of AI (cutting its guts open) toward an ancestral future: one that does not force a dichotomy between human development and environmental preservation (Krenak 2020; 2022). By distancing from the risk narrative, perspectivism, and purposiveness become two theoretical-methodological models that bring forth the visibility necessary to work with AI in a direction towards frugal and positional data models.

AI ethics through a decolonial lens: what does AI ethics look like if we take seriously the push to decolonise it?

Track: World Class Research Track

Authors: Beatrice Bonami (University of Edinburgh), Selena Nemorin (Monash University)

Presenter: Beatrice Bonami (University of Edinburgh)

Abstract: This paper offers an exploratory examination of the emerging discourse and practices on decolonising AI ethics. Whilst the field of AI ethics has made substantial progress in proposing normative frameworks for responsible innovation, these frameworks remain predominantly grounded in Western epistemological traditions, foundations which have limited the capacity of AI ethics to engage with plural ontologies and non-Western moral traditions, particularly those emerging from the Global South. Drawing from decolonial and critical data studies, this paper interrogates what an alternative, decolonial orientation to AI ethics might entail in both conceptual and practical terms. We contend that mainstream AI ethics frameworks fail to recognise epistemic diversity and to establish dialogical relations with non-Western knowledge systems, thereby reproducing colonial hierarchies of thought and governance within technological design and deployment. An answer to the question of decolonising AI ethics is not simple or straightforward and will, therefore, not be single handedly delivered in this paper, whilst we acknowledge the situatedness and limitations of our study. Whilst we understand that there are many other realms to examine for a decolonial AI ethics (such as critical infrastructure, labour, and political economy), we decide to focus on the following reflections: (1) AI as a means to redress power asymmetries and structural exploitation; (2) AI grounded in local data sovereignty and equitable epistemic participation; and (3) AI conceived as relationally entangled with humans, more-than-humans, and the Earth. These reflections foreground the material and ecological dimensions of data—its extractive infrastructures, environmental externalities, and geo-political inequalities—as central to the ethical assessment of AI. Through a literature-based exploration of these domains, we argue that decolonising AI ethics requires moving beyond the universalist and procedural tendencies of Western ethical paradigms towards pluriversal, situated, and ecologically embedded forms of ethical reasoning. The paper concludes by outlining how these reflections can serve as a theoretical scaffold for rearticulating AI ethics in and from the Global South, resisting the uncritical replication of Northern frameworks and opening pathways for epistemic plurality in global AI governance.

‘It literally feeds on data’: Co-designing Privacy Conscious and Trustworthy LLM Dialogues with End Users

Track: World Class Research Track

Authors: Adam Jenkins (King's College London), Pushpi Bagchi (University of Edinburgh), Humphrey Curtis (King's College London), William Seymour (King's College London), Mark Coté (King's College London), Jose Such (INGENIO (CSIC-Universitat Politècnica de València))

Presenter: Adam Jenkins (King's College London)

Abstract: Large Language Models (LLMs) have rapidly become ubiquitous, demonstrating remarkable proficiency in generating text and responding to prompts. Despite their potential, concerns about privacy and trust persist, yet methods for engaging end-users in designing privacy-conscious AI remain limited. This paper presents a replicable co-design methodology for engaging end-users in privacy-conscious LLM design. Through workshops with 36 participants across three speculative scenarios—mental health applications, travel assistants, and workplace assistants—we demonstrate how combining speculative scenarios with established frameworks (Grice's Maxims, Schaub's privacy design space) and trained facilitators enables meaningful participation from users without technical expertise. The co-designed dialogues reveal that users desire dynamic, context-sensitive privacy communication that leverages LLMs' conversational capabilities. We provide practitioners with a replicable methodology comprising: (1) scenario creation methods, (2) framework scaffolding approaches, and (3) facilitator training guidance, alongside design implications for privacy-conscious conversational AI.

Fast Certified Safety Checks for AI-Assisted Air Traffic Control

Track: Emerging Research

Authors: George De Ath (University of Exeter), Ben Carvell (NATS)

Presenter: George De Ath (University of Exeter)

Abstract: AI-assisted air traffic control systems need fast, explicit safety checks for proposed heading clearances: before a controller-facing tool, optimiser, or learning-based agent suggests an instruction, it must judge whether nearby aircraft remain laterally separated under uncertain speeds and turns. We present a pairwise lateral safety test under a bounded planar model in which each aircraft holds a constant speed from a known interval and may execute one fixed-rate turn. The straight-heading case reduces to an exact geometric closest-approach test; the one-turn case admits a certified time search over exact fixed-time reachable sets, giving a one-sided guarantee: any instruction it certifies safe is genuinely safe under the model. Across synthetic benchmark campaigns the implementation matches the exact analytical straight-heading rule and produces no observed false-safe outcomes against dense same-model turn-aware references, with a full decision in roughly five microseconds. It is already the pairwise safety check across several Project Bluebird air traffic control agents and evaluation suites.

Corporate sustainability as kaleidoscopes to shield AI's green economy: a critical discourse analysis on Big Tech's private reports

Track: Emerging Research

Authors: Beatrice Bonami (University of Edinburgh), Mohammad Amir Anwar (University of Edinburgh)

Presenter: Beatrice Bonami (University of Edinburgh)

Abstract: We are working with the hypothesis that "sustainability" is a kaleidoscope (cohort of mirror concepts whose reflections produce an illusion of depth/action) to overshadow Artificial Intelligence (AI) corporations' predatory activities, strategically positioned in environmental reports to divert attention from the material aspects of AI's networks of production and value chains. Sustainability is often semantically associated with preservation and longevity of organic and inorganic systems, a term which gained traction in promoting actions that could revert the societal and environmental impacts of human actions embedded by the Anthropocene. When it comes to AI, the corporate promises regarding data models often fall around the premise of AI being a green industry or technology. However, the AI industry is neither green nor sustainable if we consider the material aspect of its networks of production and global value chains. Our conceptual paper will analyze this hypothesis by performing Critical Discourse Analysis (CDA) in the keynote addresses given by Google, Palantir and Anthropic during the AI Global Summit held in India in February 2026, followed by their corporate reports launched in the weeks previous to the event affirming their plans to make AI green. Our CDA will code via NVivo the keynotes (pairing the CDA with report analysis and critical fact checking) to dissect what in the speeches can be considered "actions taken", "planned activities" and "future visions" when it comes to address the kaleidoscope concept of sustainability in either of the meanings (preservation and/or longevity). Our paper will shed light on the nuances regarding AI sustainability discourse and present discursive strategies to distract attention from exploitative and predatory practices that pave AI development worldwide.

A rapid evidence review of evaluation techniques for large language models in legal use cases: trends, gaps, and recommendations for future research

Track: World Class Research Track

Authors: Joshua Kelsall (University of Warwick), Xingwie Tan (Warwick University), Aislinn Bergin (Nottingham University), Jiahong Chen (Sheffield University), Maria Waheed (Nottingham University), Tom Sorell (Warwick University), Robert Procter (Warwick University), Maria Liakata (Queen Mary University London), Jenny Chim (Queen Mary University London), Serene Chi (Warwick University)

Abstract: The legal profession faces mounting pressures, including case backlogs and limited access to legal services. Large language models (LLMs), such as OpenAI's GPT series, have been touted as potential solutions, promising to streamline tasks such as legal drafting, summarisation, analysis, and advice. Proponents argue these models can enhance efficiency, accuracy, and access to justice. However, significant risks remain. LLMs are prone to bias, factual hallucinations, and opaque reasoning processes, which can have severe consequences in high-stakes legal contexts. For responsible use in law, legal use cases must be accurately operationalised into LLM tasks that are sensitive to legal settings, as do the evaluation metrics used to evaluate LLMs performing those tasks. This paper presents a rapid literature review of LLM research in legal contexts since ChatGPT-4's release in March 2023. We examine how legal tasks are operationalised for LLMs and what evaluation metrics are used, with a focus on how these align—or fail to align—with real-world legal practice. We argue that existing studies often overlook the institutional, organisational, and professional contexts in which these tools would be deployed. This oversight limits the practical relevance of current evaluations and proposes directions for more contextually grounded research and responsible deployment strategies.

A case study in sustainable AI: Applying carbon-reducing techniques in weather forecasting frameworks

Track: Emerging Research

Authors: Kara Moraw (EPCC, The University of Edinburgh)

Presenter: Kara Moraw (EPCC, The University of Edinburgh)

Abstract: This poster presents a case study in sustainable AI applied to weather forecasting frameworks such as ECMWF's Anemoi and NCAR's CREDIT. Machine Learning (ML) is transforming weather and climate science, increasingly complementing traditional methodologies for forecasting and analysing complex atmospheric phenomena. There is evidence that ML-based solutions can be computationally more efficient than traditional models while achieving comparable accuracy. As minimising the energy consumed by data centres and HPC resources is key to help tackle climate change, leveraging ML for scientific computing applications has the potential to significantly reduce the environmental impact of weather and climate modelling. This work aims to better understand current best practice, limitations and opportunities around fusing ML with scientific workflows to drive energy efficiency. To this end, we collate appropriate sustainable AI techniques across the community. We then quantify energy use with and without those optimisations and document their applicability in domain-specific ML frameworks. In order to examine the full ML lifecycle, we design a case study, fine-tuning ML-based weather forecasting models for extreme rain events in the UK.

Pack defender: proactive defense against packet attacks in NoCs using an XGBoost-RNN model

Track: World Class Research Track

Authors: Shengkai Hu (School of Electronics and Computer Science, University of Southampton), Haoyu Wang (Department of Engineering Science, University of Oxford), Basel Halak (School of Electronics and Computer Science, University of Southampton), BooJoong Kang (School of Electronics and Computer Science, University of Southampton)

Abstract: The Network-on-Chip (NoC) serves as the critical communication backbone in modern Multi-Processor Systems-on-Chip (MPSoCs), particularly for Deep Learning (DL) hardware where it underpins the reliable execution of machine learning models by facilitating efficient data and weight exchange. However, the NoC is vulnerable to stealthy packet-based attacks initiated by malicious Intellectual Property (IP) cores. Such attacks can severely degrade NoC latency and throughput, which are critical for efficient DL inference, and even compromise the correctness of model execution. Current detection methods are inherently reactive; they identify anomalies by monitoring global system features only after an attack has manifested, lacking the foresight to anticipate impending threats. To address this, we propose Pack Defender, a proactive NoC security framework based on temporal behavior modeling that forecasts future system states and reuses its partial prediction generative model for detection, eliminating the need for a separate module. Experimental results show strong predictive power, with average/top-three similarities of 83%/92% for Source-Level Packet Dropping (SLPD) and 90%/94% for In-Network Packet Diversion (INPD). The low Mean Absolute Error (0.05 for SLPD, 0.03 for INPD) further confirms its accuracy. The detection model (XGBoost) achieves 100% accuracy, with recall rates of 96% and 99% for SLPD and INPD respectively, significantly outperforming state-of-the-art methods lacking proactive prediction.

Accent Awareness in Open Speech LLMs: A Benchmark on British Isles English on Speech Understanding

Track: Emerging Research

Authors: Tareq Al Jaber (University of Hull)

Presenter: Aylin Naebzadeh (University of Hull)

Abstract: This paper introduces a novel benchmarking framework for evaluating state-of-the-art open-source automatic speech recognition (ASR) models and audio language models (Audio LLMs) on native speech from six regional varieties of British English: Southern England, Midlands, Northern England, Welsh, Scottish, and Irish. Unlike existing English ASR benchmarks that predominantly feature near-standard accents, this benchmark is explicitly designed to stress-test models on accent diversity and speaker demographics. It is built on a curated corpus of transcribed utterances from speakers balanced across gender and region.We propose two complementary evaluation components: (1) ASR performance assessed w.r.t. a suite of classical, embedding-based semantic, and LLM-as-Judge metrics; and (2) paralinguistic understanding, operationalized as zero-shot classification of speaker gender and regional accent. Results are broken down by gender, region, and per model, enabling fine-grained analysis of accent-specific degradation and demographic disparities. Our findings reveal substantial and persistent performance gaps across British regional accents and genders, even for leading open-source models, and demonstrate cases where conventional error metrics fail to capture semantic understanding under heavy accent variation. We release the benchmark, evaluation code, and per-accent results as an open resource to support reproducible research and fairer speech technologies.

Enhancing Cultural Heritage through AI-Driven Historical Avatars in Augmented Reality

Track: Emerging Research

Authors: Stephen Uzor (Anglia Ruskin University)

Presenter: Stephen Uzor (Anglia Ruskin University)

Abstract: This paper describes the design and evaluation of Living Exhibitions, a mobile Augmented Reality (AR) application that allows users to engage in real-time conversations with AI-driven digital avatars of historical figures. A user study (n = 17) examined engagement, trust, historical accuracy, and how people interact with the system in practice. The findings highlight the potential for AI in heritage settings, alongside user suggestions regarding authenticity, customisation, trust, and accessibility that any future deployment will need to address.

Machine Learning for Dynamic Predictions of Household Energy Demand

Track: World Class Research Track

Authors: Jenny Cox (UKRI AI Centre for Doctoral Training in AI for Sustainability, University of Southampton), Patrick James (University of Southampton), Zhanxing Zhu (University of Southampton), Thomas Rushby (University of Southampton)

Presenter: Jenny Cox (UKRI AI Centre for Doctoral Training in AI for Sustainability, University of Southampton)

Abstract: The UK Government is due to introduce the new Home Energy Model (HEM) to replace the existing Standard Assessment Procedure (SAP) for estimating the energy efficiency of homes for domestic Energy Performance Certificates (EPCs). The opensource methodology for the HEM proposes a core physical building model capable of accepting code “wrappers” that can predict home energy usage based on occupant specific information with the potential to develop an interactive EPC interface for end-users. Such an initiative would re- quire the development of models that can more accurately predict home energy usage based on inputs relating to both physical building and occupant characteristics. Using data from the Smart Energy Research Laboratory (SERL), this research combines exploratory analysis of half-hourly energy use across socioeconomic groups during the coldest week of a UK study period spanning July 2021 to June 2024 with the CatBoost machine learning algorithm for pre- dicting daily household energy demand. The model achieved RMSE values of 14.62 kWh/day within sample and 22.53 kWh/day out-of-sample, outperform- ing existing benchmarks. This work contributes to development of HEM code wrappers for dynamic EPC interfaces and provides insights to inform design of energy efficiency policy which support the reduction of carbon emissions at household level.

What does Agentic AI brings to Elderly Care?

Track: World Class Research Track

Authors: Hazrat Ali (University of Stirling, United Kingdom), Ruhul Amin Khalil (UAE University, United Arab Emirates), Kashif Ahmad (Munster Technological University, Ireland)

Presenter: Hazrat Ali (University of Stirling, United Kingdom)

Abstract: The global ageing population necessitates new and emerging strategies for caring for older adults. In this article, we explore the potential for transformation in elderly care through Agentic Artificial Intelligence (AI), powered by Large Language Models (LLMs). We discuss how Agentic AI facilitates proactive, autonomous decision-making in elderly care. Personalized tracking of health, cognitive care, and environmental management, all aimed at enhancing independence and high-level living for older adults, represents important areas of application. With the potential to significantly transform elderly care, Agentic AI also raises profound concerns about data privacy and security, decision independence, and access. We share key insights to emphasize the need for ethical safeguards, privacy protections, and transparent decision-making. Our goal in this article is to provide a balanced discussion of both the potential and the challenges of Agentic AI, and to offer insights into its responsible use in elderly care, aligning it with the requirements and vulnerabilities specific to the elderly. Finally, we identify the priorities for the academic research communities to achieve human-centred advancements and integration of Agentic AI in elderly care. To the best of our knowledge, this is one of the first comprehensive studies explicitly focused on LLM-based Agentic AI for elderly care. Hence, we address the literature gap by analyzing the unique capabilities, applications, and limitations of LLM-based Agentic AI in elderly care.

TEASBench: Tracking Evolving AI and Systems through Realistic Cost, Accuracy, and Performance Benchmarking

Track: Emerging Research

Authors: Yinsicheng Jiang (University of Edinburgh), Yufan Zhao (University of Edinburgh), Liang Cheng (University of Edinburgh), Arno Proeme (EPCC, University of Edinburgh), James Richings (EPCC, University of Edinburgh), Mark Klaisoongnoen (EPCC, University of Edinburgh), Daniyal Arshad (EPCC, University of Edinburgh), William Lucas (EPCC, University of Edinburgh), Aaron Zhao (Imperial College London), Wenda Li (University of Edinburgh), Boris Grot (University of Edinburgh), Adrian Jackson (EPCC, University of Edinburgh), Nick Brown (EPCC, University of Edinburgh), Luo Mai (University of Edinburgh), Edoardo Ponti (University of Edinburgh)

Presenter: Yinsicheng Jiang (University of Edinburgh)

Abstract: We introduce TEAS (Tracking Evolving AI and Systems), an evolving benchmark suite for the Cost, Accuracy, and Performance of AI inference on state-of-the-art workloads. Existing benchmarks such as MLPerf Inference and InferenceX target short sequences, throughput metrics, and dense architectures, missing the shift toward sparse Mixture-of-Experts (MoE), reasoning models, and agentic tool use. TEAS pairs a basic-tasks workload over four sparse-MoE models with an agentic workload over two open-weight models, run across five hardware platforms (A100, H100, H200, B200, MI355X) and two inference engines (vLLM, SGLang) for 360 experiments, with closed-form analytical models for MoE memory and bandwidth. We release TEASBench, a modular Kubernetes and shell pipeline that generates reproducible experiments from CSV, template, and config triples and surfaces results as hardware maps and CAP radar plots on a public website. We also outline extensions to non-GPU accelerators (Cerebras CS-3, Tenstorrent), multi-agent and multimodal workflows, and integrated profiling. The suite, results, and documentation are available at https://www.teasbench.com.

EEsizer: LLM-Based AI Agent for Sizing of Analog and Mixed Signal Circuit

Track: World Class Research Track

Authors: Chang Liu (The University of Edinburgh), Danial Chitnis (The University of Edinburgh)

Abstract: The design of Analog and Mixed-Signal (AMS) integrated circuits (ICs) often involves significant manual effort, especially during the transistor sizing process. While Machine Learning techniques in Electronic Design Automation (EDA) have shown promise in reducing complexity and minimizing human intervention, they still face challenges such as numerous iterations and a lack of knowledge about AMS circuit design. Recently, Large Language Models (LLMs) have demonstrated significant potential across various fields, showing a certain level of knowledge in circuit design and indicating their potential to automate the transistor sizing process. In this work, we propose EEsizer, an LLM-based AI agent that integrates large language models with circuit simulators and custom data analysis functions, enabling fully automated, closed-loop transistor sizing without relying on external knowledge. By employing prompt engineering and Chain-of-Thought reasoning, the agent iteratively explores design directions, evaluates performance, and refines solutions with minimal human intervention. We first benchmarked 8 LLMs on six basic circuits and selected three high-performing models to optimize a 20-transistor CMOS operational amplifier, targeting multiple performance metrics, including rail-to-rail operation from 180 nm to 90 nm technology nodes. Notably, OpenAI o3 successfully achieved the user-intended target at 90 nm across three different test groups, with a maximum of 20 iterations, demonstrating adaptability and robustness at advanced nodes. To assess design robustness, we manually designed a bias circuit and performed a variation analysis using Gaussian-distributed variations on transistor dimensions and threshold voltages. Overall, the results demonstrate the potential of LLMs to accelerate AMS circuit design by reducing manual effort while maintaining efficiency and adaptability across circuits, models, and technology nodes.

Towards a new paradigm of ‘datafied development

Track: World Class Research Track

Authors: Mohammad Amir Anwar (University of Edinburgh), Nanna Bonde Thylstrup (University of Copenhagen), Rafael Grohmann (University of Toronto), Julian Posada (Yale University)

Presenter: Mohammad Amir Anwar (University of Edinburgh)

Abstract: Development is broadly defined as progressive change or improvement in human well-being through multifaceted approaches, such as infrastructure enhancement, trade and investment, access to decent work, skills development, and social inclusion. It is increasingly entangled with data-centric practices and digital infrastructures like artificial intelligence systems, platforms, the internet, computers, and mobile phones. This trend warrants attention on how various digital innovation initiatives intertwine with conflicting visions of enhanced human well-being and perpetuate historically harmful power structures. As a result, this article introduces the concept of "datafied development" to critically examine human development in relation to contemporary data-intensive technologies. Datafied development allows us to be attentive to new modalities of governance: from retrospective measurement to prediction and pre-emption; from public statistics to proprietary data assets; from territorial state infrastructures to transnational cloud and platform systems; and from development expertise grounded in field-based knowledge to knowledge practices increasingly shaped by datasets, proxies, dashboards, and algorithmic outputs. In doing so this article, together with the other collections in this special issue integrates multidisciplinary scholarship on critical data studies, science and technology studies into conversation with development studies. It aims to facilitate a substantial cross-disciplinary conversation on the increasing intersections between theories and practices of development initiatives and datafication. The special issue proposes analysing datafied development across four interrelated domains: epistemic, infrastructural, economic and geopolitical.

A Privacy-Constrained Approach to Medical Data Acquisition

Track: Emerging Research

Authors: Kurt Butler (The University of Edinburgh), Edward Moroshko (The University of Edinburgh), Sotirios Tsaftaris (The University of Edinburgh)

Presenter: Kurt Butler (The University of Edinburgh)

Abstract: Acquiring medical data can be costly and time consuming, but it is necessary to produce models that generalize across patient populations and contexts. For model developers, it may be unclear which data sets are most useful to acquire to train a good model. In this work, we discuss a heuristic strategy for selecting data to acquire using only privacy-preserving summaries of what these data sets contain.

Multi-Agent Coordination of Multi-Energy Buildings

Track: Emerging Research

Authors: Samara Vazquez (University of Southampton)

Abstract: The UK's Net Zero 2050 commitment requires integrating renewable energy while managing intermittency and grid stability. Buildings consume 59% of UK electricity and offer significant flexibility through multi-energy coordination. However, realising this at scale demands coordinated multi-building systems rather than isolated optimization. Building energy management systems (BEMS) minimize costs by exploiting flexibility, but model-based approaches struggle with thermal non-linearity, while reinforcement learning offers scalable, model-free solutions. Centralised multi-building coordination faces scalability and privacy limitations; decentralised multi-agent approaches are necessary but underdeveloped. This work addresses the gap by proposing a decentralised multi-agent framework for multi-energy building communities that simultaneously ensures thermal comfort, respects individual building autonomy, and leverages multi-energy flexibility for grid support. We present a centralised optimization baseline to benchmark the emerging MARL coordination framework.

Beyond Token-Level Diversity: Context-Mediated Exploration for RL Post-Training of Language Models

Track: Emerging Research

Authors: Jim Dilkes (University of Southampton), Vahid Yazdanpanah (University of Southampton), Sebastian Stein (University of Southampton)

Abstract: RL post-training improves LLM agents only if they explore useful new behaviours, but classical action-noise exploration does not translate to token generation. We show that token-level entropy interventions fail to improve task reward, and propose Instruction-Conditioned Exploration (ICE): varying prompt instructions during training to induce strategy-shaped rather than noise-shaped behavioural diversity. On Overcooked, ICE unlocks task progress standard PPO cannot reach, albeit inconsistently, motivating a self-distillation mechanism to transfer these behaviours into the deployed policy.

EZCap: A Single-Camera Markerless Motion Capture System for Digital Health & Monitoring

Track: Emerging Research

Authors: David Hollinger (University of Essex), Behnam Gholami (University of Essex), Bernard Liew (University of Essex), Xiaojun Zhai (University of Essex)

Presenter: David Hollinger (University of Essex)

Abstract: Establishing reliable measurements during stereotypical movements is important for monitoring health over time. However, current motion capture systems are generally limited due to setup cost, time, and subject-matter expertise. To overcome these barriers, markerless motion capture using artificial intelligence (AI) has transformed the world of digital health and monitoring due to their convenience, lower cost, and increasing reliability (Lam et al., 2023). These systems have become increasingly popular in clinical practice (SOURCE) and sport environments (SOURCE). In this study, we present EZCap, a video-based AI motion capture system for doing human movement analysis. The purpose of this study was to validate EZCap against validated against gold-standard laboratory system (e.g., Vicon) fifteen participants (Sex: 10 Males, 5 Females, age: 30.27yrs ± 9.63yrs, height: 1.74m ± 0.10m, weight: 73.65kg ± 11.69kg), during sit-to-stand (STS) and countermovement jump (CMJ) tasks. Our results indicate EZCap as a reliable and accurate approach to digital health and monitoring across clinical and sport-related movements.

Multimodal Longitudinal Lung Function Pattern Analysis to Predict Asthma and COPD

Track: Emerging Research

Authors: Ammarah Farooq (King's College London), Minghua Zheng (King's College London), Elif Dogu (King's College London), Claire Steves (King's College London), Mona Bafadhel (King's College London)

Presenter: Ammarah Farooq (King's College London)

Abstract: Asthma and chronic obstructive pulmonary disease (COPD) are major causes of mortality and healthcare expenditure in the UK. Diagnosis is often delayed, with disease progression recognised only after substantial and irreversible lung function decline. Current clinical assessment relies largely on cross-sectional spirometry thresholds, which provide limited insight into the long-term dynamics of lung health. This ongoing project aims to develop a data-driven framework for modelling lung function trajectories across adulthood using longitudinal data from the TwinsUK cohort. The analysis will identify distinct lung function patterns associated with asthma and COPD. Leveraging the twin study design, the research will examine the relative contributions of genetic, environmental, and behavioural factors to disease onset and progression. The expected outcome is a predictive framework for earlier identification of individuals at risk of accelerated lung function decline. This work aims to support improved risk stratification, earlier intervention, and more targeted prevention strategies for respiratory diseases.

LABIIUM: AI-Enhanced Zero-configuration Measurement Automation System

Track: World Class Research Track

Authors: Emmanuel Olowe (The University of Edinburgh), Danial Chitnis (The University of Edinburgh)

Presenter: Emmanuel Olowe (The University of Edinburgh)

Abstract: The complexity of laboratory environments requires solutions that simplify instrument interaction and enhance measurement automation. Traditional tools often require configuration, software, and programming skills, creating barriers to productivity. Previous approaches, including dedicated software suites and custom scripts, frequently fall short in providing user-friendly solutions that align with programming practices. We present LABIIUM, an AI-enhanced, zero-configuration measurement automation system designed to streamline experimental workflows and improve user productivity. LABIIUM integrates an AI assistant powered by Large Language Models (LLMs) to generate code. LABIIUM's Lab-Automation-Measurement Bridges (LAMBs) enable seamless instrument connectivity using standard tools such as VSCode and Python, eliminating setup overhead. To demonstrate its capabilities, we conducted experiments involving the measurement of the parametric transfer curve of a simple two-transistor inverting amplifier with a current source load. The AI assistant was evaluated using different prompt scenarios and compared with multiple models, including Claude Sonnet 3.5, Gemini Pro 1.5, and GPT-4o. An expert solution implementing the Gradient-Weighted Adaptive Stochastic Sampling (GWASS) method was used as a baseline. The solutions generated by the AI assistant were compared with the expert solution and a uniform linear sweep baseline with 10,000 points. The graph results show that the LLMs were able to successfully complete the most basic uniform sweep, but LLMs were unable to develop adaptive sweeping algorithms to compete with GWASS. The evaluation underscores LABIIUM's ability to enhance laboratory productivity and support digital transformation in research and industry, and emphasizes the future work required to improve LLM performance in Electronic Measurement Science Tasks.

Policy-Governed LLM Routing with Intent Matching for Instrument Laboratories

Track: World Class Research Track

Authors: Emmanuel Olowe (The University of Edinburgh), Danial Chitnis (The University of Edinburgh)

Presenter: Emmanuel Olowe (The University of Edinburgh)

Abstract: AI tutoring systems in engineering labs face a tension between providing sufficient assistance and preserving learning opportunities. Existing systems typically offer instructors limited control over assistance timing, content, or cost. This paper describes a routing and governance system for LLM-based lab assistance comprising two components: Routiium, an OpenAI-compatible gateway that manages multiple LLM backends with configurable prompt modifications and usage logging, and EduRouter, a policy-aware routing service that enforces per-lab budgets, approval workflows, and embedding-based question matching. We evaluated the system using trace-driven simulation calibrated from two engineering labs (LED characterization, RC circuit analysis) and a 100-query replay through live models. In simulations, governed policies (P1/P2) increased challenge-alignment index from 0.90 to 0.98 and overlay-adherence score from 0.69 to 0.87 compared to ungoverned operation (P0). The productive-struggle window metric increased from 1.4 to 3.6 simulated turns before high-scaffold hints appeared. In the 100-query replay, EduRouter routed 75% of queries to a local model, reducing token costs by 66% ($0.087 vs. $0.26 for all-premium routing) while maintaining canonical hit rate of 1.0 for the curated 89-intent question bank. We release Routiium, EduRouter, canonical-task tooling, and simulator configurations to support replication and future classroom studies.

BEACON-M: Continuous Many-Objective Benchmark Problems with Explicit, Adjustable Correlation

Track: Emerging Research

Authors: Sam Tebbet (University of Exeter), George De Ath (University of Exeter), Tinkle Chugh (University of Exeter)

Presenter: Sam Tebbet (University of Exeter)

Abstract: Multi-objective optimisation is prevalent in artificial intelligence research, arising across settings including hyperparameter optimisation to reinforcement learning. Correlations between objectives are natural in such real-world problems: they fundamentally shape the landscapes of multi-objective optimisation problems and influence the search behaviour of a multi-objective optimiser. Yet existing continuous multi-objective benchmarks rarely provide mechanisms to define them explicitly. We present BEACON-M, a continuous multi-objective benchmark generator that constructs smooth objectives via Gaussian Process Random Fourier Features and offers explicit user-defined correlations. We provide two distinct structured correlation matrix formats: single-group and two-group structures with ensured reproducibility through instances governed by analytically derived validity constraints. To demonstrate the benchmark's use, we run experiments with state-of-the-art multi-objective evolutionary algorithms across different numbers of decision variables, objectives and correlation coefficients. Performance depends on the interplay between problem dimensionality, in both objective and decision spaces, and the correlation structure and strength, with dominance-based methods strongest under aligned objectives and low decision dimensionality, and decomposition methods more robust under two-group conflict at higher decision dimensionality. These findings demonstrate that such problems can support the development of effective correlation-aware algorithms that can tackle complex real-world optimisation challenges.

EEspice: A Modular Circuit Simulation Platform with Parallel Device Model Evaluation via Graph Coloring

Track: World Class Research Track

Authors: Xuanhao Bao (The University of Edinburgh), Danial Chitnis (The University of Edinburgh)

Abstract: As modern analogue/mixed-signal design increasingly relies on optimization-in-the-loop flows, such as AI and LLM-based sizing agents that repeatedly invoke SPICE—efficient, accurate high-performance simulators have become an indispensable foundation for modern integrated circuit (IC) design. However, the computational cost of evaluating nonlinear models, particularly for BSIM models, remains a significant bottleneck. In standard parallelization approaches, devices such as transistors are easily distributed across processors. The subsequent stamping phase, where each device's contributions are added to the shared system matrix, often creates a bottleneck. Because multiple processor cores compete to update the same matrix elements simultaneously, the system is forced to process tasks one at a time to avoid errors. This paper introduces EEspice, an open-source circuit simulation framework whose modular architecture decouples device model evaluation into independently replaceable kernels, enabling a parallel stamping strategy that overcomes this bottleneck. It partitions MOSFET instances into independent color groups, which can be processed in parallel. Our results show that on a 64-core workstation, the proposed approach achieves up to 45x speedup over single-thread performance when conflicts are low. Our analysis also explores how performance depends on circuit topology.

A Spatio-Temporal Graph Similarity Approach to Characterising Strategic Variation Among Air Traffic Control Trainees

Track: Emerging Research

Authors: Mudathir Ayomide Ekundayo (University of Exeter, UK), Richard Everson (University of Exeter, UK), George De Ath (University of Exeter, UK)

Presenter: Mudathir Ayomide Ekundayo (University of Exeter, UK)

Abstract: This paper presents an initial investigation into a data-driven, spatio-temporal approach to characterising strategic variation among Air Traffic Control (ATC) trainees. Current assessment frameworks, including that used by the National Air Traffic Services (NATS) remain largely outcome-focused, evaluating whether a trainee avoided conflicts and demonstrated the required competencies, rather than examining the underlying strategies through which traffic was managed. To address this gap, each trainee's behaviour is represented as a time-evolving sequence of nearest-neighbour adjacency matrices encoding the lateral and vertical configuration of aircraft within the sector at each timestep. Pairwise dissimilarity between trainee sequences is computed using Open-End Dynamic Time Warping (OE-DTW) and the Hamming distance, and the resulting dissimilarity matrix is projected into a low-dimensional space via Multidimensional Scaling (MDS) to reveal strategic groupings among trainees. The approach is evaluated on 58 trainees completing an exercise from the NATS College training curriculum . The MDS projection reveals a dominant cluster of strategically similar trainees alongside three interpretable outliers, whose lateral and vertical flight profiles exhibit meaningful and identifiable differences relative to the cluster. These findings suggest that the proposed framework has potential as a complementary tool to conventional outcome-based ATC assessment, offering instructors a richer, behaviour-informed perspective on trainee performance.

Automation and the Ecology of Expertise in Air Traffic Control

Track: Emerging Research

Authors: Edmond Awad (University of Exeter), Robert Procter (University of Warwick)

Abstract: AI-enabled automation is entering air traffic control, but its workforce effects cannot be understood only by asking which tasks are automated. This paper argues that automation in air traffic control must also be evaluated by how it reshapes practice, control-loop engagement, and responsibility for rare high-consequence events. We introduce five ATCO--AI role archetypes and discuss how these archetypes affect expertise, coordination, and governance.

Explaining Failures of Cyber-Physical Systems with Actual Causality

Track: World Class Research Track

Authors: Khen Elimelech (King's College London), Tom Yaacov (King's College London), David A. Kelly (King's College London), Hana Chockler (King's College London), Moshe Y. Vardi (Rice University)

Presenter: Tom Yaacov (King's College London)

Abstract: Modern autonomous Cyber-Physical Systems (CPSs), such as self-driving cars, face increasingly complex demands, and yet are expected to act reliably. The black-box nature often characterizing such systems, especially those relying on neural components, makes it impossible to fully verify the system behavior prior to deployment. Unfortunately, unexpected failures—cases when the system does not comply with its specification—are inevitable and may have catastrophic implications. To improve trust in the system and facilitate future mitigation after a failure occurs, it is important to try to derive an explanation for the unexpected system behavior. This paper introduces the novel concept of leveraging the framework of actual causality for CPS failure explanation. Up until now, this framework was only used to derive explanations in the context of simple systems, such as image classifiers. This paper addresses the theoretical gaps and provides the guidance needed to allow for correct explanation derivation in the CPS domain. Beyond the theoretical contribution, the paper presents two novel, practical, system-agnostic explanation derivation algorithms, allowing to prioritize either explanation optimality or derivation efficiency. The approach is demonstrated and evaluated in the context of a neural-network-controlled autonomous car, designed to avoid collisions.

Trust in Air Traffic Control: Implications for Agent-Based Systems

Track: Emerging Research

Authors: Robert Procter (University of Warwick), Mark Rouncefield (University of Siegen)

Presenter: Robert Procter (University of Warwick)

Abstract: Air traffic control (ATC), the safe, orderly and efficient management of planes through controlled airspace is a complex task. In this article, we explore the roles that different kinds of trust play in ATC. Our study provides the essential foundations for examining the conditions that need to be satisfied if Air Traffic Controllers are to establish and sustain trust relationships with technical innovations, such as agent-based systems, and so ensure their safe and dependable embedding within ATC work.

In the Moment of Glitch: Engaging with Misalignments in Ethical Practice

Track: World Class Research Track

Authors: Rachal Garrett (KTH Royal Institute of Technology, Sweden), Kat Hawkins (Centre for Dance Research, Coventry University), Pat Brundell (University of Nottingham), Simon Cast;e-Green (University of Nottingham), Paul Tennent (University of Nottingham), Feng Zhou (University of Nottingham), Airi Lampinen (KTH Royal Institute of Technology Stockholm, Sweden), Kristina Höök (KTH Royal Institute of Technology Stockholm, Sweden), STEVE BENFORD (University of Nottingham)

Abstract: Summary for UKAIRS: This paper reflects on an ethical matter that arose during an artist-led exploration of dancers with limb differences engaging in contact improvisation with a robot arm. The dancers repeatedly encountered a self-protective behaviour in the robot in which it froze if pushed to its limits of safe movement, which they tended to do. This was initially described by the team as being a “glitch”. However, the use of this term caused an ethical debate about the dancers’ own experience of disability in which they were often viewed as being glitchy - as problems in need of technical solutions. In response, the paper surfaces the concept of ‘glitch’ as site of ethical exploration and debate as part of a more responsible approach to synergistic human interaction with robots an AI. The original paper was published at the 2025 ACM Computer-Human Interaction (CHI) conference. The research was funded by the UKRI Trustworthy Autonomous Systems Hub. The following is the abstract from the original paper: “Glitches – moments when technologies do not work as desired – will become increasingly common as industrially-designed robots move into complex contexts. Taking glitches to be potential sites of critical ethical reflection, we examine a glitch that occurred in the context of a collaborative research project where professional dancers with different disabilities improvised with a robotic arm. Through a first-person account, we analyse how the dancer, the robot, and the rest of the research team enacted ethics in the moment of glitch. Through this analysis, we discovered a deep and implicit ethical misalignment wherein our enactments of ethics in response to the glitch did not align with the values of the project. This prompted a critical re-engagement with our research process through which we forged a dialogue between different ethical perspectives that acted as an invitation to bring us back into ethical alignment with the project’s values.”

Diagnosing Lower Extremity Arteriovenous Diseases Using Agentic LLMs

Track: World Class Research Track

Authors: Zicen Liao (Queen Mary University of London)

Abstract: Large language models (LLMs) achieve strong results on medical benchmarks, but real out-patient diagnosis is iterative and requires sta-ble, guideline-consistent reasoning rather than one-shot predictions. We study this problem in lower-extremity arteriovenous (LEA) dis-eases and introduce LEA-Dialog, a multi-turn diagnostic dialogue dataset covering six com-mon diseases, with stage annotations for each turn and guideline-grounded probability trends. Building on this dataset, we further propose a process-aligned agentic framework for real-world patient consultations. In this framework, the patient interacts with a Doctor agent, while a Teacher agent audits guideline adherence and reasoning quality, provides corrective feedback to reffne the Doctor agent's behaviour, and ul-timately produces a structured diagnostic re-port. An Evaluator agent then derives both turn-level and ffnal rewards from the diagnos-tic conversations between the patient and the Doctor agent for PPO-based policy optimisa-tion, while carefully selected high-quality di-agnostic conversations are further used for su-pervised ffne-tuning. For training, the patient is also simulated by a Patient agent that emu-lates real patient responses. In addition to ffnal diagnostic accuracy, we evaluate unsupported claims, trend correctness, diagnostic-tree sim-ilarity, and repeated-run stability with a new metric, Rate-n. Experiments on both online and offffine LLMs show that process-aligned multi-turn diagnosis reduces reasoning drift, improves stability, and yields the largest gains for smaller offffine models.

Learning Temporal Embeddings for Multi-level Clustering of Physiological States in PICU

Track: Emerging Research

Authors: Brian O'Donovan (School of Engineering, University of Edinburgh), Hollan Haule (School of Engineering, University of Edinburgh), Patricia A. Jones (Child Life and Health University of Edinburgh, UK), Ian Piper (Usher Institute University of Edinburgh, UK), Tsz-Yan Milly Lo (Paediatric Critical Care Unit RHCYP Edinburgh, UK Usher Institute University of Edinburgh, UK), Laura Moss (Department of Clinical Physics and Bioengineering NHS Greater Glasgow and Clyde, UK School of Medicine University of Glasgow, UK), Javier Escudero (School of Engineering, University of Edinburgh)

Presenter: Brian O'Donovan (School of Engineering, University of Edinburgh)

Abstract: Early physiological dynamics from the first 24 hours of monitoring in the paediatric intensive care unit (PICU) may contain clinically important information, but remain difficult to summarise. We propose a self-supervised pipeline for learning temporal embeddings of short physiological windows from routinely collected PICU data. Windows from the first 24 hours of each patient's stay are encoded with a joint contrastive and reconstruction objective. Embeddings are clustered into physiological states, and patient-level representations are obtained from the state trajectories. The learned embeddings produce separable physiological states and reveal distinct patient subgroups with interpretable phenotypes in a PICU cohort with clinical-grade data, suggesting that temporal representation learning can support unsupervised phenotyping in intensive care.

AI-Enabled Optimization of SRAM Cell Design Using the Async-SADEA Framework

Track: Emerging Research

Authors: Vihar Georgiev (University of Glasgow, UK), Naveen Kumar (University of Glasgow, UK), Tongbao Yang (University of Glasgow, UK), Bo Liu (University of Glasgow, UK), Luiz Felipe Aguinsky (University of Glasgow, UK), Ankit Dixit (University of Glasgow, UK)

Abstract: This study demonstrates a Async-SADEA based machine learning framework for optimizing the static noise margin (SNM) of Static Random Access Memory (SRAM) cells. A conventional six-transistor (6T) SRAM cell is employed to establish the proof of concept. The work investigates the optimization of individual transistor-level parameters to achieve near-ideal figures of merit for inverter circuits, with particular emphasis on the voltage transfer characteristics (VTC), using the SADEA framework. By applying the proposed machine learning algorithm, critical device parameters including gate length, height, doping concentration, and area factors are systematically optimized to enhance overall SRAM cell performance.

FedORA: Federated Objective-Resolved Adaptation for Non-Stationary Environments

Track: World Class Research Track

Authors: Wai Fong Tam (Queen Mary University of London), Songyuan Li (Queen Mary University of London), Ahmed M Abdelmoniem (Queen Mary University of London)

Presenter: Wai Fong Tam (Queen Mary University of London)

Abstract: Federated Learning (FL) at the edge faces a critical challenge: client data distributions drift over time due to changing environments, user behavior, or operational contexts. Existing drift-adaptive FL methods are \emph{objective-agnostic}, assuming all clients share identical adaptation priorities. In practice, edge clients may prioritize rapid adaptation to current conditions (short-horizon) or sustained performance across historical scenarios (long-horizon). We propose \textbf{FedORA} (Federated Objective-Resolved Adaptation), an edge-cloud framework that routes clients to adaptation strategies based on both diagnosed drift type and their declared objectives. FedORA uses multi-signal drift detection and objective-aware grouped aggregation to prevent conflicting model updates from diluting each other's performance. We introduce dual-objective metrics—Average Current Local Accuracy (ACLA) for short-term adaptation and Average Accuracy over Past Tasks (AAoPT) for long-term oriented clients.

Human Proxies for Accelerated Learning in Human-Machine Teaming Tasks

Track: Emerging Research

Authors: Ash Simmons (University of Exeter), George De Ath (University of Exeter), Richard Everson (University of Exeter)

Presenter: Ash Simmons (University of Exeter)

Abstract: Effective human-machine teaming requires autonomous agents capable of coordinating with unpredictable human partners. Traditional reinforcement learning methods relying on live human feedback or extensive behavioural data are often prohibitively expensive, while purely simulated approaches, such as population play, tend to overfit to artificial conventions and fail under Zero-Shot Coordination with real human partners. To bridge this gap, a novel framework which trains teaming agents against synthetic “human proxies” is proposed. Evaluated within a modified version of Gymnasium’s Lunar Lander environment, these proxies are explicitly designed to emulate genuine cognitive diversity, operating under reaction delays and perceptual limitations analogous to human players. Training a Recurrent PPO agent against a diverse pool of these proxies allows it to integrate sequential observations and account for human suboptimalities, leading to an agent more familiar with the choices made by real human partners. Preliminary trials indicate that proxy-trained agents achieve higher mean rewards, exhibit greater robustness to human error, and are perceived as easier to collaborate with compared to self-play and population-play baselines.

Late Breaking Results: Automated Design of Device Fingerprinting Circuits for FPGAs

Track: Emerging Research

Authors: James Moore (Queen's University Belfast)

Abstract: Designing device fingerprinting circuits for FPGAs is challenging due to limited knowledge about device layout and manufacturing- induced biases, often resulting in poor uniqueness and biased responses. Existing mitigation techniques are typically device-specific, time-consuming, and difficult to generalize. We propose an auto- mated EDA tool, that analyses the target FPGA to optimize circuit layout to maximise extracted entropy and mitigate systematic bias. Across 100-device testbed, the tool outperformed comparable manual designs, achieving near-ideal uniqueness and bias. These results demonstrate that automated device-aware design can reliably estimate and approach the best achievable performance on a given FPGA

Large Language Models: Are They Ready for Data Science?

Track: Emerging Research

Authors: Ozan Evkaya (University of Edinburgh), Wenqi Ni (University of Edinburgh), Miguel de Carvalho (University of Edinburgh)

Abstract: A key goal of this project is to evaluate the performance of the OpenAI Advanced Data Analytics (ADA) tool on DS tasks (e.g., data wrangling, exploratory analyses, fitting models) For that purpose, OpenAI Code Interpreter assistant will be used for the specific tasks repeatedly by mimicking the idea of Monte Carlo experiments borrowed from the field of Statistics. By incorporating the Code Interpreter within GPT assistant, using zero-shot style prompts, the project aims to measure the performance of this tool numerically alongside the automation of DS discussions. To give a solid measure of the performance, certain aspects of the generated outcome (i.e. correctness, text similarity, verbosity) will be mainly focused on. In the literature, experiments done so far are mainly limited to a certain number of queries mostly and not resembling the idea of repeated sampling, necessary for the uncertainty measurement on top of the generated responses from GPT's ADA extension. To overcome this issue, by following different strategies (different data set file upload, specific Data Science related prompts and more advanced statistical concepts explanation), the project will capture a comprehensive overview of the certain behavior of the DA capabilities of GPT models. For the tool development, OpenAI platform instructions have been considered to create a new DA-capable GPT within a personal execution environment. Using the curated tasks from the specific parts of the four DS quadrants mentioned in the work of \cite{DeBie2022}, the project aims to contribute the discussions about the automation of the Data Science using empirical Gen-AI based results.

AI Agents Can Pass the IUCN Red List Exam – At What Cost?

Track: Emerging Research

Authors: Shane Weisz (University of Cambridge), Michael Dales (University of Cambridge), Sadiq Jaffer (University of Cambridge), Anil Madhavapeddy (University of Cambridge)

Presenter: Shane Weisz (University of Cambridge)

Abstract: This emerging research forms part of a new collaboration between the University of Cambridge's Department of Computer Science and the International Union for Conservation of Nature (IUCN), investigating how agentic AI can support species extinction risk assessments for the IUCN Red List of Threatened Species. Our early results show that frontier AI agents can now pass the official IUCN Red List assessor-training exam – but only since Claude 4.5 Sonnet's release in September 2025. The Red List exam is a mandatory certification required by assessors to verify they have the skills and knowledge to assess species extinction risk according to the official Red List categories and criteria – passing it shows that AI meets a capability floor ahead of consideration for real-world Red List use. Work currently in progress includes deeper model/harness comparisons, and an analysis of the cost tradeoffs of agentic AI (financial, environmental, reputational, and opportunity costs) – particularly relevant for a conservation NGO like the IUCN. Next steps will then test real-world assessment workflows such as consistency checking of draft assessments. We have an accompanying live demo: an interactive web app showing Claude Code scoring >90% in under 5 minutes on the 25-question, 3-hour exam by spinning up agents in parallel, calling tools, and answering with citations to official sources.

An LLM-based Transistor Level Layout Generation

Track: Emerging Research

Authors: YUAN MENG (The University of Edinburgh), Danial Chitnis (The University of Edinburgh)

Presenter: YUAN MENG (The University of Edinburgh)

Abstract: The back-end design stage is critical in transistor-level integrated circuit design, requiring substantial manual effort and affecting area efficiency, design rule compliance, and performance. Conventional layout automation often depends on large proprietary datasets, limiting generalization across circuit structures. In this work, we propose an LLM-based framework for generating complex standard-cell layouts with limited metal layers. The framework combines LLM reasoning, connectivity-aware placement, dynamic programming-based routing, and a DRC feedback loop for iterative area compression. Experiments on five SkyWater 130 nm PDK standard cells show that GPT-5.2 achieves competitive layout area, with a MUX cell using 93\% of the area of human-optimized implementations. These results demonstrate the feasibility of LLM-based agents for transistor-level layout generation.

Quantifying Opportunities of Reducing CO2 Footprint of Dairy Farms with Different On-Site Renewable Generation and Regional CO2 Grid Mix

Track: Emerging Research

Authors: John Callaghan (University of Strathclyde), Jack Mackenzie (University of Strathclyde), Lina Stankovic (University of Strathclyde), Vladimir Stankovic (University of Strathclyde)

Presenter: John Callaghan (University of Strathclyde)

Abstract: Dairy farms have been increasing their reliance on machinery to resulting in a heavier reliance on the United Kingdom's (UK) energy grid supply which utilises non-renewable sources as part of the grid mix , emitting predominantly carbon dioxide (CO2) [DESNZ REF], thus increasing the carbon footprint of dairy farms by proxy [Carbon Intensity REF]. As farmers have to comply with future net zero policies, it is important for them to develop on-site renewable energy, leaving them with a question of which renewable is most appropriate for their farm. This is currently not answered by current models such as Cool Farm Tool (CFT) as it is used for already established energy sources to estimate greenhouse gas (GHG) emissions based on the year. This paper proposes a data-driven methodology for estimating more localised carbon emissions from dairy farm electricity import metrics and electricity demand, including the contribution of on-site renewables generation. This is mapped against regional half-hourly carbon intensity data from the Carbon Intensity API. An unsupervised learning approach is used to cluster daily electricity demand into representative operating profiles, each assigned a profile-specific carbon emission factor. Results indicate that regional and profile-based carbon accounting can shift estimated electricity-related carbon output by 34.5%, highlighting the need for more precise carbon models as agriculture prepares for future net zero policy requirements.

Towards Diverse Models: An Investigation of Shortcut Learning

Track: Emerging Research

Authors: Nehal Yasin (University of Southampton), Jonathon Hare (University of Southampton), Antonia Marcu (University of Southampton)

Abstract: Neural networks remain brittle under distribution shifts, largely because models fail to learn diverse predictive patterns from the training data. One key reason for this is referred to as shortcut learning, where models latch onto one predictive pattern and ignore other predictive patterns which might help generalise to unseen settings. Despite substantial research effort, shortcut learning remains an open problem. We argue this is because the field has advanced by introducing mitigation strategies that are built upon foundational assumptions that have never been rigorously evaluated. We systematically examine these assumptions and find they are context-sensitive, holding under narrow conditions but failing to generalise. Our results have direct implications for how the field evaluates and addresses shortcut learning. Based on our findings, we demonstrate that the current framing of shortcut learning requires systematic analysis and clear definitions. We believe that the objective of diversification, rather than a shortcut suppression, could be a more fruitful research direction.

An activity schedules calibration framework for mobility simulation models

Track: Emerging Research

Authors: Yahya Gamal (AI for Collective Intelligence, University of Glasgow), Tatsuya Mitomi (AI for Collective Intelligence, University of Glasgow, UK), Esra Suel (Urban Analytics, Department of Geography, University of Zurich, Switzerland), Gary Polhill (Information & Computational Sciences Department, The James Hutton Institute, UK), Alison Heppenstall (AI for Collective Intelligence, University of Glasgow)

Presenter: Yahya Gamal (AI for Collective Intelligence, University of Glasgow)

Abstract: Multi-agent mobility simulations, particularly agent-based models (ABMs), are complex. They include multiple interacting actors with each other and with geospatial road networks. This making it challenging to calibrate the components of such models to reality, particularly daily activity schedules. Mobility models follow two approaches for activity scheduling calibration: (1) fixed activity sequences with calibrated trip parameters such as trip times; and (2) restructured activity sequences for calibrating trip parameters. The former ignores potential uncertainties with the initial allocation of activities, and the latter can deviate from empirical data during the restructuring of activities. To address this, we propose an activity calibration framework allowing for optimal reallocation of activities across a population without disrupting the empirical activities distributions. We showcase the framework using our mobility ABM `MATraM' in the context of Tillydrone, Aberdeen, Scotland. Initial results indicate that the proposed framework minimises errors in simulated trip times without disrupting activity distributions.

Uncovering Clinical Structure in Autoregressive Health Foundation Models

Track: Emerging Research

Authors: Inigo Stibbe (Imperial College London), Padmanabhan Ramnarayan (Imperial College London), Aldo Faisal (Imperial College London)

Abstract: Health Foundation Models are trained on vast amounts of Elec- tronic Health Record data to then be used for a range of down- stream tasks such as predicting patient risk, synthetic data gener- ation, and supporting clinical decision-making. For these types of models to be implemented into a healthcare system, more work needs to be done to assess what the model is learning and con- vince stakeholders of its safety. This paper investigates model rep- resentations at the point of ICU admission via qualitative UMAP plots and linear probing for mortality prediction. We find that the model representation of the token at ICU admission clusters pa- tients very well in comparison to count vector representation. In addition, training a linear probe for 30-day mortality prediction on the token embedding gives a 14.7% increase in PR-AUC score compared to the baseline.

When Does AI Consensus Matter? The Effect of AI Consensus on Advice Adherence in a Modified Trust Game

Track: Emerging Research

Authors: Thom Griffith (Cardiff University), Phillip Morgan (Cardiff University)

Abstract: Artificial intelligence systems are increasingly deployed as advisors to human decision-makers, yet little is known about how people respond to recommendations produced by groups of AI agents. Drawing on theories of social influence, this study investigates whether consensus among AI advisors affects advice adherence and whether this effect depends on how advisors are presented. Participants completed a modified Trust Game in which they received recommendations from a panel of five AI advisors. Consensus was manipulated by varying the proportion of advisors recommending the same action. Advisors were presented either as intentional agents with individual identities and first-person language or as machine-generated outputs. The primary outcome measure was advice adherence. Preliminary analyses indicate that the influence of consensus depends on advisor presentation. When advisors were framed as intentional agents, adherence increased with increasing levels of consensus. In contrast, when identical recommendations were presented as machine outputs, consensus had little effect on participant behaviour. These findings suggest that agreement among AI systems does not automatically function as a social influence cue, but instead becomes influential when users interpret AI advisors as intentional entities. This work contributes to understanding social influence in human–AI interaction and has implications for the design of multi-agent AI decision-support systems.

Synthetic Dataset Creation as Governance Infrastructure for Embodied AI Bias Evaluation

Track: Emerging Research

Authors: Afifah Kashif (University of Cambridge)

Abstract: The integration of foundation models into robotic systems extends the bias risks documented in conversational AI into the physical and interactional space of human-robot interaction, where existing content-based evaluation methods are structurally insufficient. Recent work has called for certification-grade risk assessment for embodied AI, but the evaluation infrastructure required to operationalize that certification does not yet exist and cannot be collected from naturally occurring deployment. Safety-critical scenarios cannot be ethically staged at scale, intersectional demographic coverage is structurally sparse, and refusal patterns and behavioral channels are invisible to content-based methods. This paper proposes synthetic dataset creation, extending a methodology validated by the author in the conversational setting, as the operational mechanism through which anticipatory HRI governance can be implemented. Five construction moves are translated from the conversational to the embodied setting, with implications for evaluation practice and certification policy.

Deep Learning-Based Individual Re-identification of Reef Fish: A Component Variation Study

Track: Emerging Research

Authors: Tedi Yankov (University of Oxford), Oliver King (University of Oxford), Niki Amini-Naieni (University of Oxford), Andrew Zisserman (University of Oxford), Cait Newport (University of Oxford)

Presenter: Tedi Yankov (University of Oxford)

Abstract: Reliable individual re-identification of reef fish in the wild is a key bottleneck in automated marine ecology. We present a component variation study on a Picasso triggerfish re-identification pipeline, systematically comparing five detection / segmentation methods (RT-DETR, YOLO11-seg, Cascade Mask R-CNN, Mask2Former, SAM), four feature extraction backbones (MiewID, MegaDescriptor-L, TransReID, DINOv2), and three metric learning objectives (softmax, triplet, ArcFace) across 60 pipeline combinations evaluated with 3-fold cross-validation on a controlled lab subset. Our best-performing configuration according to performance and latency - MiewID with RT-DETR and softmax loss (composite score 0.962) - substantially outperforms alternatives. Contrary to our expectations, mask cropping does not improve re-identification; background context preserved by bounding box crops provides discriminative signal in cluttered underwater scenes.

Adaptive Sampling for Hydrodynamic Stability

Track: Emerging Research

Authors: Anshima Singh (University of Manchester), David J Silvester (University of Manchester)

Presenter: Anshima Singh (University of Manchester)

Abstract: We present an adaptive, data-efficient method for detecting stability boundaries in parametrized fluid flow problems. Our approach couples a neural network classifier with KRnet, a flow-based deep generative model, to automatically focus expensive computational fluid dynamics simulations in regions of highest uncertainty. Starting from a small uniform set of simulations, the method iteratively refines its sampling distribution using Shannon entropy as an uncertainty measure, progressively resolving bifurcation boundaries with far fewer high-fidelity solves than conventional parameter sweeps. We demonstrate the approach on a differentially heated cavity problem, where each simulation takes approximately four hours, showing that three adaptive steps are sufficient to resolve a complex, non-monotone stability boundary across a wide range of fluids. The framework is scalable and transferable to any physical system where locating critical transitions is computationally expensive.

Skipping the IME? Keyboard-Friendly Multilingual Prompting for LLM Agents

Track: Emerging Research

Authors: Xiaoxiao Zhouyi (University of Bristol), Yanzhen Li (University of Bristol)

Abstract: Many multilingual users do not compose prompts only in polished native script. Everyday interaction often passes through an input method editor (IME), transliteration, abbreviation, code-switching, or mixed native/Latin technical vocabulary. This emerging research asks whether LLM agents remain useful when users skip final IME conversion and type keyboard-friendly forms instead. We introduce ime-eval, an agent-native evaluation scaffold that pairs native-script prompts with keyboard variants across Chinese pinyin and other multilingual input styles. Preliminary zh-CN evidence from Codex and a small Claude Code check shows that clean pinyin can preserve quality on routine prompts, but aggressive initials and toneless pinyin on homophone-sensitive content produce sharp regressions. A mixed workflow, where ordinary instruction is typed in pinyin but exact high-risk terms remain in Hanzi, restores correctness on targeted homophone tasks. The work contributes a reproducible evaluation artifact and a practical design rule for multilingual human-agent interaction: keyboard-friendly prompting may be viable for routine scaffolding, but exact names, domain terms, and homophones need native-script anchoring.

Knowledge Repositories and Reasoning Engines: Decoupling Monolithic AI Systems

Track: Emerging Research

Authors: Yihong Chen (University of Oxford & GenAI Hub), Joshua Yi Ren (University of Oxford & GenAI Hub), Yarin Gal (University of Oxford & GenAI Hub)

Presenter: Yihong Chen (University of Oxford & GenAI Hub)

Abstract: Intelligence is not merely a system's performance within a single world frozen in time, but the ca pacity to generalize in an ever-changing world, varying across time and viewpoints. As worlds shift, knowledge changes with the world, whereas reasoning persists across worlds. Modern AI systems, however, couple knowledge and reasoning, binding behavior to training distributions and failing to distinguish world-specific truths from general reasoning. This coupling leads to practical strains on efficiency, reliability, and adaptability. This whitepaper proposes decoupling knowledge and reasoning through Com-Putare, a conceptual framework in which a reusable reasoning engine operates over world-specific knowledge repositories. It outlines the principles required to construct decoupled intel ligent systems and provides criteria for measuring their degree of knowledge–reasoning decoupling. This perspective shifts truth authority from centralized model owners to user- or community-controlled knowledge, redefining AI as systems that respect explicit knowledge boundaries and remain portable across worlds, an important step toward trustworthy and user-sovereign AI.

Enhancing Logical Reasoning in Language Models via Symbolically-Guided Monte Carlo Process Supervision

Track: World Class Research Track

Authors: Xingwei Tan (University of Sheffield), Macro Valentino (University of Sheffield), Mahmud Elahi Akhter (Queen Mary University of London), Maria Liakata (Queen Mary University of London), Nikolaos Aletras (University of Sheffield)

Presenter: Xingwei Tan (University of Sheffield)

Abstract: Large language models (LLMs) have shown strong performance in many reasoning benchmarks. However, recent studies have pointed to memorization, rather than generalization, as one of the leading causes for such performance. LLMs, in fact, are susceptible to content variations, demonstrating a lack of robust planning or symbolic abstractions supporting their reasoning process. To improve reliability, many attempts have been made to combine LLMs with symbolic methods. Nevertheless, existing approaches fail to effectively leverage symbolic representations due to the challenges involved in developing reliable and scalable verification mechanisms. In this paper, we propose to overcome such limitations by synthesizing high-quality symbolic reasoning trajectories with stepwise pseudo-labels at scale via Monte Carlo estimation. A Process Reward Model (PRM) can be efficiently trained based on the synthesized data and then used to select more symbolic trajectories. The trajectories are then employed with Direct Preference Optimization (DPO) and Supervised Fine-Tuning (SFT) to improve logical reasoning and generalization. Our results on benchmarks (i.e., FOLIO and LogicAsker) show the effectiveness of the proposed method with gains on frontier and open-weight models. Moreover, additional experiments on claim verification data reveal that fine-tuning on the generated symbolic reasoning trajectories enhances out-of-domain generalizability, suggesting the potential impact of the proposed method in enhancing planning and logical reasoning.

Fundamental Reasoning Paradigms Induce Out-of-Domain Generalization in Language Models

Track: World Class Research Track

Authors: Mingzi Cao (University of Sheffield), Xingwei Tan (University of Sheffield), Mahmud Elahi Akhter (Queen Mary University of London), Macro Valentino (University of Sheffield), Maria Liakata (Queen Mary University of London), Xi Wang (University of Sheffield), Nikolaos Aletras (University of Sheffield)

Presenter: Xingwei Tan (University of Sheffield)

Abstract: Deduction, induction, and abduction are fundamental reasoning paradigms, core for human logical thinking. Although improving Large Language Model (LLM) reasoning has attracted significant research efforts, the extent to which the fundamental paradigms induce generalization has yet to be systematically explored. In this study, we shed light on how the interplay between these core paradigms influences LLMs' reasoning behavior. To this end, we first collect a new dataset of reasoning trajectories from symbolic tasks, each targeting one of the three fundamental paradigms, to abstract from concrete world knowledge. Then, we investigate effective ways for inducing these skills into LLMs. We experiment with a battery of methods including simple fine-tuning, and more complex approaches to increase model depth, or transform a dense model to a mixture-of-experts. We comprehensively evaluate induced models on realistic out-of-domain tasks, that are entirely formulated in natural language and contain real-world knowledge. Our results reveal that our approach yields strong generalizability with substantial performance gains (up to 14.60) across realistic tasks.

Project Bluebird Demo

Track: Emerging Research

Authors: Richard Everson (University of Exeter & Alan Turing Institute)

Presenter: Richard Everson (University of Exeter & Alan Turing Institute)

Abstract: This is a proposal to hold an interactive demonstration of the Project Bluebird digital twin of airspace, an interactive air traffic control game and an AI agent performing the same task as participants.

CellTypeAI: cell annotation for scRNA-seq using local generative-AI

Track: World Class Research Track

Authors: Rufus Daw (AI-Hub in Generative Models, University of Manchester), Harry Deijnen (University of Manchester), Magnus Rattray (AI-Hub in Generative Models, University of Manchester), John Grainger (University of Manchester)

Presenter: Rufus Daw (AI-Hub in Generative Models, University of Manchester)

Abstract: Single-cell RNA sequencing (scRNA-seq) cell annotation techniques rely on the matching of known defining marker genes to a given cell population. However, these methods may lack robustness to dynamic fluctuations in cell marker expression between patients, samples and pathologies. The advent of easy-to-implement predictive technologies, like generative-AI (gen- AI), has facilitated the introduction of computational workflows that improve otherwise inaccurate context-dependent cell type annotation. Here, we introduce CellTypeAI, a streamlined, scalable program developed for tissue context-dependent cell annotation of scRNA-seq datasets using modern gen-AI models, enhanced by retrieval augmented generation methods. Our implementation builds upon local gen-AI hosting technologies and directly integrates into scRNA-seq analysis pipelines. We show that CellTypeAI provides improved annotation accuracy compared to current conventional annotation methods and nascent cloud-based gen-AI approaches. As CellTypeAI leverages locally-run AI models, it can be applied to sensitive datasets, unlike approaches utilising online gen-AI tools such as ChatGPT, DeepSeek, or Claude. CellTypeAI presents a novel solution for tissue-specific cell type identification, overcoming traditional marker-based limitations via locally-deployed gen-AI models.

Transformed Latent Variable Multi-Output Gaussian Processes

Track: World Class Research Track

Authors: Xiaoyu Jiang (University of Manchester), Xinxing Shi (University of Manchester), Sokratia Georgaka (University of Manchester), Magnus Rattray (University of Manchester), Mauricio Alvarez Lopez (University of Manchester)

Presenter: Xiaoyu Jiang (University of Manchester)

Abstract: Multi-Output Gaussian Processes (MOGPs) provide a principled probabilistic framework for modelling correlated outputs but face scalability bottlenecks when applied to datasets with high-dimensional output spaces. To maintain tractability, existing methods typically resort to restrictive assumptions, such as employing low-rank or sum-of-separable kernels, which can limit expressiveness. We propose the Transformed Latent Variable MOGP (T-LVMOGP), a novel framework that scales MOGPs to a massive number of outputs while preserving the capacity to capture meaningful inter-output dependencies. TLVMOGP constructs a flexible multi-output deep kernel by mapping inputs and output-specific latent variables into an embedding space using a Lipschitz-regularised neural network. Combined with stochastic variational inference, our model effectively scales to high-dimensional output settings. Across diverse benchmarks, including climate modelling with over 10, 000 outputs and zero-inflated spatial transcriptomics data, TLVMOGP outperforms baselines in both predictive accuracy and computational efficiency.

What does explainability, auditability and transparency mean to frontline health and social care practitioners? A co-design study for safer human-AI collaboration in NHS Continuing Healthcare

Track: Emerging Research

Authors: Dwynwen Hopcroft (University of Lancashire)

Presenter: Dwynwen Hopcroft (University of Lancashire)

Abstract: This poster submission presents an early-stage co-design study exploring what explainability, auditability and transparency mean to frontline nurses and social workers involved in NHS Continuing Healthcare decision-making. The study addresses a gap between technical definitions of trustworthy AI and the practical needs of professionals working in high-stakes, evidence-based care settings. Using workshops and a simple AI demonstrator, the research will generate practitioner-informed design guidance for safer and more effective human-AI collaboration.

Designing Victim Personas for Anti-Social Behaviour Practitioner Training: A Multi-Factorial LLM Embodiment Framework

Track: Emerging Research

Authors: Rachel Hill (Swansea University), Tom Owen (Swansea University), Julian Hough (Swansea University), Muneeb Ahmad (Swansea University)

Presenter: Rachel Hill (Swansea University)

Abstract: Artificial Intelligence (AI) is increasingly adopted within UK public sector governance, yet technological innovation within Anti-Social Behaviour (ASB) intervention and practitioner training remains limited. This work presents the Coded Persona Embodiment Framework (CPEF), a multi-factorial framework built using existing research on LLM embodied tone, ASB forms, and psychological theory, designed to create complex LLM persona embodiment for practitioner training. The work further presents a pilot study that tested Claude's Haiku 4.5 model's ability to embody a synthetic ASB victim persona presented comparatively in prose versus CPEF's code. Findings indicated that Claude's Haiku 4.5 model was able to understand and embody CPEF, but iterative prompt and system engineering was required to improve conversational pacing, immersion, and contextual appropriateness, demonstrated through interaction excerpts. CPEF opens the possibility for complex persona embodiment in ASB victim simulation training settings.

Band Gap Prediction of Two-Dimensional Materials Using a Gradient-Boosted Feature Selection Approach

Track: World Class Research Track

Authors: Ben Rowlinson (University of Edinburgh), Ram Ramamoorthy (University of Edinburgh), Jacqueline M. Cole (University of Cambridge), Themis Prodromakis (University of Edinburgh)

Presenter: Ben Rowlinson (University of Edinburgh)

Abstract: Two-dimensional (2D) inorganic crystals are a class of materials that are gaining significant attention for use in electronic and optoelectronic devices. Among many other exciting applications, 2D materials offer a range of beneficial properties to such devices owing to charge carrier confinement, high carrier mobility, tunable band gaps, strong light–matter interactions, and atomically thin geometries that enable excellent electrostatic control and mechanical flexibility. In parallel, data-driven approaches to predictions of inorganic material properties have gained considerable attention as computationally lightweight surrogate models for properties of interest. This is particularly important when screening candidate materials for particular sets of structure–property relationships. Many of these approaches have targeted three-dimensional bulk crystalline materials. In this work, we develop a set of data-driven models for predicting the properties of 2D layered, van der Waals, and ultrathin film materials, namely, thermodynamic stability, metallicity, and electronic band gap. We train the models on materials sourced from open-source computational databases of 2D materials (Alexandria_2D, C2DB, MC2D, and 2DMatpedia) and use chemically relevant elemental, physical, and compositional features as input. The large feature space is reduced to a subset of critical features by a statistical and gradient-boosted feature selection strategy. The models are fully interpretable with feature relevance scores and SHapley Additive exPlanations analysis assessing the global and local influence of feature values. The classifiers for thermodynamic stability and metallicity achieve F1-scores of 0.832 and 0.870 and accuracies of 89.7% and 89.7%, respectively. The regressor model for the band gap achieves an R2 of 0.883, a mean-absolute error (MAE) of 0.317 eV, and a root-mean-squared error (RMSE) of 0.485 eV on the in-distribution test set. We assess the band gap predictor regressor against a 2D material band gap data set (N ∈ 177) manually extracted from the academic literature to quantify the model's ability to predict outside of the training distribution, achieving an R2 of 0.334, an MAE of 0.675 eV, and an RMSE of 0.961

eV. These results demonstrate the efficacy of feature selection in producing fully explainable machine learning surrogate models for high-throughput property prediction for 2D materials.

I’m Sorry Driver, I’m Afraid I Can’t Do That: Appraising the Safety of LLMs within Automotive Contexts

Track: World Class Research Track

Authors: Kim Littler (UKRI AI Centre for Doctoral Training in Safe Artificial Intelligence Systems (University of York)), Shaun Feakins (UKRI AI Centre for Doctoral Training in Safe Artificial Intelligence Systems (University of York)), Ibrahim Habli (UKRI AI Centre for Doctoral Training in Safe Artificial Intelligence Systems (University of York)), Robert Palin (Jaguar Land Rover (JLR))

Abstract: This paper appraises recent frameworks within AI development to integrate LLMs into control tasks in automotive contexts from the perspective of safety assurance. This work has built upon the rapid integration of LLMs across automotive settings. However, we find that at present, these frameworks face significant challenges, limiting their efficacy in real-time safety-critical contexts. Firstly, we consider conceptual challenges, including the fact that deployers are faced with a dual challenge, wherein they must assure a model which has been developed upstream, i.e. as general-purpose tools by the large AI labs, in a downstream context, i.e. into specific vehicle architectures. Secondly, we consider concrete challenges from across existing standards. We show that there are currently both fundamental engineering constraints covered in ISO21448, such as latency, and novel LLM-specific issues, such as alignment-related issues covered in ISO/PAS8800. We ground both examples in a concrete introductory, experimental case study exploring an existing open-source repository, Talk2Drive. We present a safety argument in order to make explicit the limitations of existing solutions. Nonetheless, given that the use of LLMs in automotive contexts is being explored at a technical level and operationalised, we propose potential assurance mechanisms for LLM-related hazardous events going forward.

From Automation to Authority: Introducing CCALS and REACT-HX, a Human-Expertise Framework for Safeguarding Agency in AI-Mediated Societies

Track: World Class Research Track

Authors: Michal Cholerzynski (Northumbria University), David Hastings (Northumbria University)

Presenter: Michal Cholerzynski (Northumbria University)

Abstract: This exploratory paper examines how the growing integration of artificial intelligence (AI) into workplaces, platforms, and public institutions reshapes the distribution of expertise, institutional agency, and human participation within socio-technical systems. While existing research on AI and labour markets highlights productivity gains and relatively modest net employment effects, such perspectives often overlook how AI infrastructures can redistribute discretion, accountability, and decision-making authority even when employment remains stable. Drawing on literature from labour economics, CSCW, HCI, organisational studies, and AI governance, we argue that tools introduced as assistive technologies can progressively assume greater authority in coordinating work and shaping institutional decisions. To address this shift, we introduce REACT-HX, a conceptual framework identifying five structural dimensions for sustaining meaningful human involvement in AI-mediated environments: Role Reflexivity (RR), Employment Resilience (ER), Agency Preservation (AP), Citizen-Impact Awareness (CIA), and Technology-Human Synergy (THS). We also propose Citizen-Centred AI Labour Studies (CCALS), a research agenda that reframes AI-mediated work through the lens of citizenship, emphasising that workers, IT professionals, managers, and service recipients retain the right to understand, contest, and oversee institutional decision-making. Together, REACT-HX and CCALS provide a foundation for analysing and designing AI-mediated systems that preserve human agency, professional expertise, accountability, and institutional legitimacy.

Early Explorations at the Intersection of Clay, AI, and Embodied Making

Track: Emerging Research

Authors: Allison Waxberg (Northumbria University)

Abstract: This poster presents a pilot study exploring the relationship be- tween clay and AI, as both a making material and a design method. Drawing on aspects of research through design, clay and a short set of prompts were distributed to participants to engage with in their own time and space. They articulated their experience with clay in relation to AI: asking AI what to make, comparing AI-suggested objects with self-directed ones, and exploring whether there are meaningful parallels between working with clay and working with AI. Early findings suggest themes around AI and legibility, the gap between AI pattern recognition and human making intention, and broader themes around posture, attention and clay as emotional regulator. The work invites a perspective around Creativity for Change, offering some early insights in to how embodied making practices together with AI can open new possibilities for creative participation.

Quantifying Explainable AI-introduced signal noise on ECG data with Spectral Entropy

Track: World Class Research Track

Authors: Nathan Blake (King's College London), David A. Kelly (King's College London)

Presenter: Nathan Blake (King's College London)

Abstract: Explainability techniques are used to assess the output of various deep learning models. This is especially true in healthcare, where models need to be trusted and decisions justified. Explainability (XAI) tools use heuristics which often add signal noise to the explanation “core”. It is not always obvious what is signal from the model and what is noise from the XAI. We propose the use of spectral entropy as a measure of noise in XAI output. We demonstrate its usefulness in the context of classifying arrhythmias in an ECG dataset with different post hoc explainability techniques.

A kind of magic? Artificial Intelligence-assisted recording: developing principles for an AI-informed social work curriculum

Track: World Class Research Track

Authors: Richard Ingram (University of Dundee)

Presenter: Richard Ingram (University of Dundee)

Abstract: The deployment of AI tools in social work practice settings isadvancing at a pace that social work education has yet to match.Drawing on baseline findings from an independent evaluation ofMagic Note, an AI-assisted recording tool deployed acrossa Scottish Local Authority social work department, this paper usesthe empirical experiences of a workforce encountering AI in prac-tice as a window into the underdeveloped implications for qualify-ing education. The evaluation, which involved 152 practitionerssurveyed, seven in focus groups, and over 255 free-text responses,surfaced five core tensions: administrative relief versus the reflec-tive value of writing; AI assistance versus professional judgment;efficiency versus relational practice; accessibility and inclusion ver-sus deskilling risk; and enthusiasm versus ethical complexity. Thepaper develops each tension into substantive educational implica-tions and proposes five curriculum principles for social work pro-grammes to engage with AI proactively rather than reactively. Theargument is that students need to arrive in practice alreadyequipped with the critical, ethically grounded capacity to engagewith AI as agents rather than subjects.

When autonomy breaks: the hidden existential risk of AI

Track: World Class Research Track

Authors: Joshua Krook (University of Southampton)

Presenter: Joshua Krook (University of Southampton)

Abstract: AI risks are typically framed around physical threats to humanity, a loss of control or an accidental error causing humanity's extinction. However, I argue in line with the gradual disempowerment thesis, that there is an underappreciated risk in the slow and irrevocable decline of human autonomy. As AI starts to outcompete humans in various areas of life, a tipping point will be reached where it no longer makes sense to rely on human decision-making, critical thinking or even creativity. What may follow is a process of gradual de-skilling, where we lose skills that we currently take for granted. Traditionally, it is argued that AI will gain human skills over time and that these skills are innate and immutable in humans. By contrast, I argue that humans may lose such skills as critical thinking, decision-making and even creativity in an AGI world. The biggest threat to humanity is, therefore, not that machines will become more like humans, but that humans will become more like machines.

Functional Alignment Can Mislead: Examining Model Stitching

Track: World Class Research Track

Authors: Damian Smith (University of Southampton), Harvey Mannering (University of Southampton), Antonia Marcu (University of Southampton)

Presenter: Damian Smith (University of Southampton)

Abstract: With the widespread adoption of AI, a better understanding of neural networks is becoming a priority. Researchers can measure how well neural networks perform a task, but would also like to understand how different networks perform the task. A useful tool that might help shed light on the inner workings of neural networks would be one that meaningfully determines when two networks are different. In this ICML 2025 Spotlight work, we focus on a technique called model stitching. Informally, stitching strongly focuses on what decision a network makes. We argue that this focus can abstract away from how the network makes that decision. As a result, stitching can consider two networks to be similar if they make the same decision but based on very different evidence (or rules). This becomes problematic since not all rules are equally likely to hold when networks are deployed in the wild. Thus, networks can have qualitatively different performance in the real world but can falsely be considered similar if compared using approaches such as stitching. In our paper we create a series of carefully designed experiments to illustrate this and encourage the deep learning community to consider such settings when evaluating other network comparison tools.

Leveraging variational autoencoders for multiple data imputation

Track: World Class Research Track

Authors: Sara Wade (University of Edinburgh)

Presenter: Sara Wade (University of Edinburgh)

Abstract: Missing data persists as a major barrier to data analysis across numerous applications. Recently, deep generative models have been used for imputation of missing data, motivated by their ability to learn complex and non-linear relationships. In this work, we investigate the ability of variational autoencoders (VAEs) to account for uncertainty in missing data through multiple imputation. We find that VAEs provide poor empirical coverage of missing data, with underestimation and overconfident imputations. To overcome this, we investigate generalisation of VAEs, developed from a Bayesian framework, to provide calibrated uncertainty and robustness to model misspecification. We apply the approach to transcriptomics datasets with various simulated missingness scenarios. Finally, we show that single imputation in transcriptomic data can cause false discoveries in downstream tasks and employing multiple imputation with generalised VAEs can effectively mitigate these inaccuracies.

Signatures of Electron–Hole Hopping in Myoglobin Peroxidase Activity Revealed by Deep Mutational Learning

Track: World Class Research Track

Authors: Alperen Dalkiran (School of Informatics, University of Edinburgh), Christoph Küng (Institute of Physical Chemistry,Department of Chemistry, University of Basel), Rosario Vanella (Institute of Physical Chemistry, Department of Chemistry, University of Basel), Diego A. Oyarzún (School of Informatics, University of Edinburgh), Michael A. Nash (nstitute of Physical Chemistry, Department of Chemistry, University of Basel)

Presenter: Alperen Dalkiran (School of Informatics, University of Edinburgh)

Abstract: In addition to storing molecular oxygen, myoglobin catalyzes peroxidase-like reactions involving high-valency iron(IV)-oxo species that support one-electron oxidations on a range of substrates at an open active site. In select metalloenzymes, longrange electron transfer can be mediated by hole-hopping pathways composed of aromatic residues that act as relay stations for oxidative equivalents. However, it remains unclear how sequence variations could introduce or alter such catalytic mechanisms in myoglobin. Here, we used enzyme proximity sequencing (EP-Seq) to measure the peroxidase activity levels of >6,000 human myoglobin variants. The resulting fitness landscape reveals how aromatic substitutions, in particular surface-exposed tryptophans, can enhance peroxidase activity. Using protein language models in tandem with feedforward neural networks, we trained an accurate fitness predictor on the experimental data set and applied it to screen >4 M double mutant variants. The predictions suggested a beneficial role for electron−hole-hopping mutations in improving peroxidase activity. We experimentally tested 20 high-scoring variants in a yeast display assay, all of which outperformed wild-type myoglobin. Three selected variants were also tested in soluble format and similarly showed improved performance. A focused combinatorial library yielded a top double tryptophan variant (Q92W/F107W) with 4.9-fold higher catalytic efficiency than wild type. These results show that deep mutational learning can identify myoglobin variants with enhanced peroxidase activity that are consistent with the involvement of hole-hopping pathways, with broad implications for biocatalyst and redox enzyme design.

Geometric Decoupling: Diagnosing the Structural Instability of Latent

Track: World Class Research Track

Authors: Yuanbang Liang (cardiff University; Gen AI Hub), Zhengwen Chen (cardiff University), Yu-Kun Lai (cardiff University; Gen AI Hub)

Presenter: Yuanbang Liang (cardiff University; Gen AI Hub)

Abstract: Latent Diffusion Models (LDMs) achieve high-fidelity synthesis but suffer from latent space brittleness, causing discontinuous semantic jumps during editing. We introduce a Riemannian framework to diagnose this instability by analyzing the generative Jacobian, decomposing geometry into Local Scaling (capacity) and Local Complexity (curvature). Our study uncovers a "Geometric Decoupling": while curvature in normal generation functionally encodes image detail, Out-of-Distribution (OOD) generation exhibits a functional decoupling where extreme curvature is wasted on unstable semantic boundaries rather than perceptible details. This geometric misallocation identifies "Geometric Hotspots" as the structural root of instability, providing a robust intrinsic metric for diagnosing generative reliability

Large Language Models as Nondeterministic Causal Models

Track: World Class Research Track

Paper status: Accepted

Authors: Sander Beckers (University College London)

Presenter: Sander Beckers (University College London)

Abstract: Recent work has developed, for the first time, a method for generating counterfactuals of probabilistic Large Language Models. Such counterfactuals tell us what would – or might – have been the output of an LLM if some factual prompt x had been x* instead. The ability to generate such counterfactuals is an important necessary step towards explaining, evaluating, and eventually improving, the behavior of LLMs. I argue, however, that the existing method rests on an ambiguous interpretation of LLMs: it does not interpret LLMs literally, for the method involves the assumption that one can change the implementation of an LLM's sampling process without changing the LLM itself, nor does it interpret LLMs as intended, for the method involves explicitly representing a nondeterministic LLM as a deterministic causal model. I here present a much simpler method for generating counterfactuals that is based on an LLM's intended interpretation by representing it as a nondeterministic causal model instead. The advantage of this simpler method is that it is directly applicable to any black-box LLM without modification, as it is agnostic to any implementation details. The advantage of the existing method, on the other hand, is that it directly implements the generation of a specific type of counterfactuals that is useful for certain purposes, but not for others. I clarify how both methods relate by offering a theoretical foundation for reasoning about counterfactuals in LLMs based on their intended semantics, thereby laying the groundwork for novel application-specific methods for generating counterfactuals.

AI in Smart Buildings and Construction 4.0: Implementation areas and influencing factors in construction organizations

Track: World Class Research Track

Authors: Qian Zhang (Durham University), Lu Wang (University of York)

Abstract: The construction sector faces persistent challenges in productivity, innovation, and fragmented organizational practices. Artificial Intelligence (AI) offers transformative potential in Smart Buildings and Construction 4.0 (SBC4.0), yet organizational-level implementation remains underexplored. This paper develops an integrated framework grounded in institutional theory, the resource-based view, and dynamic capabilities, validated through a survey of large Chinese construction firms using partial least squares structural equation modeling. Findings identify three key AI implementation areas — smart building operation and health, construction and material optimization, and offsite manufacturing and automation — each shaped by distinct configurations of normative and mimetic pressures, coercive forces, organizational management, technological feasibility, and resources and capabilities. The framework provides a roadmap for policymakers in industry and government navigating the intricacies of AI in SBC4.0 to enhance operational efficiency and organizational performance. It also establishes a foundation for future cross-national research on AI implementation across varied construction and regulatory contexts.

Decoupled Conformal Optimisation: Efficient Prediction Sets via Independent Tuning and Calibration

Track: World Class Research Track

Authors: Fanyi Wu (University of Manchester), Lihua Niu (University of Manchester), Samuel Kaski (University of Manchester), Michele Caprio (University of Manchester)

Abstract: Bayesian conformal optimisation methods often use the same held-out data both to search for efficientpredictionsetsandtocertifycoverageorrisk. Thiscouplingisnaturalforhigh-probability risk-control guarantees, but it is not necessary when the target is standard finite-sample marginal conformal coverage. We propose Decoupled Conformal Optimisation (DCO), a train-tune- calibrate design principle that uses an independent tuning split for efficiency-oriented structural selection and a fresh calibration split for the final conformal quantile. Conditional on the tuned structure, standard split-conformal exchangeability yields finite-sample marginal coverage for any candidate class, without a confidence parameter or multiple-testing correction. DCO therefore targets a different finite-sample guarantee from PAC-style methods: marginal conformal coverage rather than high-probability risk control. Under consistency assumptions on the coupled risk bound, the two approaches nevertheless converge to the same population threshold. Across classification and regression benchmarks, including ImageNet-A, CIFAR-100, Diabetes, California Housing, and Concrete, DCO tracks the nominal coverage level closely while often reducing average prediction-set size or interval width relative to PAC-style calibration. On ImageNet-A, for example, the average set size decreases from 26.52 to 25.26 and the 95th-percentile set size from 58.95 to 53.73; on Diabetes, the average interval width decreases from 2.098 to 1.914.

What are data centres and how sustainable are they?

Track: World Class Research Track

Authors: Alexander Williams (Queen Mary University of London), Simon Brawley (POST, UK Parliament)

Presenter: Alexander Williams (Queen Mary University of London)

Abstract: Data centres are buildings that contain the information technology equipment that most of the digital world relies on. They underpin most aspects of modern life, including economic activity, public service delivery, and how we interact with one another. They are crucial for cloud computing and artificial intelligence (AI). This research briefing, known as a POSTnote, is aimed at UK policy makers and discusses what data centres are, their presence in the UK, their operations, and their impact on different aspects of sustainability including resource usage, emissions, and local concerns. POSTnotes are impartial, accurate, and peer-reviewed briefings tailored for UK parliamentarians, produced by the Parliamentary Office of Science and Technology (POST). POST is an impartial research and knowledge exchange service based in the UK Parliament. POST works to ensure cutting-edge research evidence and expertise is available to members of both houses of parliament (House of Commons and House of Lords), covering emerging and complex science and social science topics. This briefing was produced in consultation with experts and stakeholders from academia, industry, government, and beyond.

Decision Curve Analysis Needs Prediction Under Intervention

Track: Emerging Research

Authors: Pedro Henrique da Costa Avelar (Univesity of Manchester), Alexander Pate (University of Manchester), Matthew Sperrin (University of Manchester)

Presenter: Pedro Henrique da Costa Avelar (Univesity of Manchester)

Abstract: Decision Curve Analysis (DCA) and Net Benefit (NB) are widely used to evaluate the clinical utility of prediction models across decision thresholds; however, standard implementations implicitly assume predicted risks correspond to untreated outcomes, creating a mismatch when models are used to guide interventions, biasing estimated utility. We address this limitation by integrating recent advances in counterfactual prediction evaluation into DCA. Using a simulation study, we provide results using the Conditional Loss (CL) approach to estimate counterfactual net benefit under intervention, and compare it to naïve NB estimates, showing that the naïve approach produces systematically biased results. These findings highlight the importance of accounting for intervention effects in DCA and suggest that counterfactual evaluation provides a more appropriate framework for assessing clinical prediction models intended for decision-making.

Vulnerability Analysis of Transformer-based Optical Character Recognition to Adversarial Attacks

Track: World Class Research Track

Authors: Lucas Beerens (University of Edinburgh), Des Higham (University of Edinburgh)

Presenter: Des Higham (University of Edinburgh)

Abstract: We present a novel framework to assess the resilience of state-of-the-art Transformer-based Optical Character Recognition (TrOCR) models. In this way, we develop new untargeted and targeted attack algorithms. On a benchmark handwriting data set we show that adversarial perturbations to the input reveal dramatic vulnerabilities without being noticeable to the eye.

LaMDA: A Language Model-Assisted Electronic Design Automation Framework

Track: Emerging Research

Authors: Cristian Sestito (The University of Edinburgh), Panagiota Kontou (The University of Edinburgh), Pratibha Verma (The University of Edinburgh), Atish Dixit (The University of Edinburgh), Alexandros D. Keros (The University of Edinburgh), Michael O'Boyle (The University of Edinburgh), Christos-Savvas Bouganis (Imperial College London), Themis Prodromakis (The University of Edinburgh)

Abstract: We recently introduced LaMDA, a flexible framework for language model-assisted electronic design automation. In this live demonstration, we highlight LaMDA's core capabilities, showing how it unifies large language models with commercial EDA tools—such as Cadence Virtuoso Spectre, Keysight ADS, and AMD Vivado—to streamline chip design across analogue, digital, and radio-frequency flows. Specifically, we present three representative workflows: designing operational amplifiers, creating and tuning microstrip patch antennas, and implementing FPGA modules (such as finite-state machines and matrix multiplication units).

Gen-TAS: Generative AI for Task Allocation in FPGA-GPP Heterogeneous Systems

Track: Emerging Research

Authors: Mary Kong (University of Edinburgh), Yuqin Zhao (University of Edinburgh), Semih Vazgecen (University of Edinburgh), Eleni Tselepi (University of Edinburgh), Cristian Sestito (University of Edinburgh), Themis Prodromakis (University of Edinburgh)

Abstract: Gen-TAS (Generative AI for Task Allocation Strategies) is a proposed framework that applies Retrieval-Augmented Generation (RAG) to support heterogeneous task allocation in FPGA-GPP computing systems. Determining which functions should execute in software or hardware remains a challenging design-space exploration problem that typically requires extensive profiling and optimisation. The framework analyses historical implementation data, user-defined performance objectives, and HLS design descriptions to generate FPGA-GPP allocation strategies and I/O mappings. A backend then produces outputs required for hardware deployment, enabling a workflow that combines AI-assisted decision-making with automated implementation generation. By extending recent advances in AI-assisted Electronic Design Automation towards system-level resource allocation, Gen-TAS aims to reduce development effort, accelerate architecture design, and improve the accessibility of heterogeneous FPGA-based system design.

Collaborative Learning and Inference for Multiclass Tasks

Track: Emerging Research

Authors: Hollan Haule (University of Edinburgh), Javier Escudero (University of Edinburgh)

Presenter: Hollan Haule (University of Edinburgh)

Abstract: In healthcare, classification systems must be both accurate, to improve patient outcomes and support clinical decision making, and interpretable, to align with evidence-based medicine. In previous work, we introduced the Collaborative Learning and Inference (CLaI) framework, which naturally incorporates case-based reasoning and addresses class imbalance in classification problems. Here, we extend CLaI to handle multi-class tasks typical of real-world scenarios. We evaluate the extended framework on image datasets, CIFAR10 and PathMNIST, using a lightweight embedding model, achieving strong performance with meaningful class clustering.

Why do VAEs not give high quality generations?

Track: Emerging Research

Authors: David Barber (University College London)

Abstract: We explore the origin of the well-known ‘blur’ in generations from a model trained with a Gaussian-q Variational Autoencoder (VAE). Our conclusion is that poor generation quality results from an insufficiently rich posterior approximation. We demonstrate this effect by essentially computing an exact posterior for a simple model, showing that the exact posterior is highly non-Gaussian and has sharp boundaries of responsibility in latent space.

Calibrated by Design: Toward Uncertainty-Aware Federated Prompt-Tuning for Trustworthy Low-Resource NLP

Track: Emerging Research

Authors: Ubaid Azam (University of Southampton), Shoaib Jameel (University of Southampton)

Presenter: Ubaid Azam (University of Southampton)

Abstract: Deploying NLP in safety-critical, low-resource domains such as law, healthcare, and under-represented languages requires models that can both quantify their own uncertainty and train without centralising sensitive data. Current methods deliver these properties in isolation: fully Bayesian models produce calibrated uncertainty but assume centralised data, while federated and parameter-efficient methods preserve privacy but emit over-confident point predictions. This early-stage work unifies the two. Building on two of our peer-reviewed systems, BAKER (a Bayesian deep Gaussian Process, Brier Score 0.022) and FLiP (federated lightweight prompt-tuning, 84% fewer trainable parameters), we propose distributional soft prompts: federated prompt parameters treated as probability distributions aggregated by predictive precision rather than dataset size. We pose three open questions on whether calibration survives federated aggregation and outline the experimental design to answer them.

Designing and Demonstrating a Survivor-Informed AI Prototype for Domestic Abuse Training

Track: Emerging Research

Authors: Dwynwen Hopcroft (University of Lancashire), Kelly Bracewell (University of Lancashire), Nicola Farrelly (University of Lancashire)

Presenter: Dwynwen Hopcroft (University of Lancashire)

Abstract: This submission presents a survivor-informed AI prototype for domestic abuse training. The prototype combines composite survivor narratives with approved professional guidance to support reflective, scenario-based learning. Grounded in trauma-informed and responsible AI principles, it includes safeguards for consent, contributor protection, and bounded system use. The work demonstrates how AI can support, rather than replace, professional judgement and lived-experience-informed learning in a sensitive safeguarding context.

GAIN: Generative AI Literacy and Knowledge Framework – the EDI Perspective

Track: Emerging Research

Authors: Laurence Brooks (University of Sheffield), Sara Cannizzaro (University of Sheffield), Pamela Abbott (University of Sheffield), Andiswa Mfengu (University of Cape Town), Reggie Raju (University of Cape Town), Sharon Wagg (University of Sheffield)

Presenter: Laurence Brooks (University of Sheffield)

Abstract: This study presents the results of a review of AI literacy guidelines from a select number of British and South African Universities. Current research on GenAI guidelines in higher education focuses on those produced by research-intensive institutions only. This approach does not recognise the insights that universities not classified as such, and less wealthy or historically disadvantaged universities, could make to an EDI-framed understanding of AI literacy. The objective of this study was to conduct an initial exploration of AI literacy guidelines from an equity, diversity and inclusion (EDI) perspective across a wide spectrum of institutions in the two aforementioned contexts (UK and South Africa), establishing a foundation for a more systematic and inclusive integration of these principles within broader HEI GenAI policy guidelines. This study provides some observations about distribution patterns of the guidelines across the UK and South Africa, and an outline of the extent to which GenAI literacy training frameworks appear to have incorporated principles around EDI. The conclusion of the current study is that the current distribution gap means disadvantaged institutions are missing the opportunity to shape the global debate on AI literacy.

Live Demo: Non-Frontal Face Recognition Using GANs and Memristor-Based Classifiers

Track: Emerging Research

Authors: Semih Vazgecen (University of Edinburgh), Cristian Sestito (University of Edinburgh), Spyros Stathopoulos (University of Edinburgh), Themis Prodromakis (University of Edinburgh)

Abstract: Face recognition systems have advanced significantly through deep learning techniques, delivering high performance and robustness in complex scenarios. However, these approaches incur substantial computational overhead, limiting their in situ applicability in resource-constrained platforms such as drones, where they can address challenges including non-frontal facial imagery. Memristor-based neuromorphic systems have emerged as a compelling approach for edge AI applications, combining biologically inspired processing with efficient and scalable computation. In this live demonstration, we present a face recognition framework that addresses non-frontal pose variations by integrating lightweight generative adversarial network (GAN)-based pose frontalisation with memristor-based neuromorphic recognition.

Of Mice and Mates: Automated Classification and Modelling of Mouse Behaviour in Groups using a Single Model across Cages

Track: World Class Research Track

Authors: Michael P. J. Camilleri (University of Edinburgh), Rasneer S. Bains (Mary Lyon Centre, MRC Harwell), Christopher K. I. Williams (University of Edinburgh)

Presenter: Michael P. J. Camilleri (University of Edinburgh)

Abstract: Behavioural experiments often happen in specialised arenas, but this may confound the analysis. To address this issue, we provide tools to study mice in the home-cage environment, equipping biologists with the possibility to capture the temporal aspect of the individual's behaviour and model the interaction and interdependence between cage-mates with minimal human intervention. Our main contribution is the novel Global Behaviour Model (GBM) which summarises the joint behaviour of groups of mice across cages, using a permutation matrix to match the mouse identities in each cage to the model. In support of the above, we also (a) developed the Activity Labelling Module (ALM) to automatically classify mouse behaviour from video, and (b) released two datasets, ABODe for training behaviour classiers and IMADGE for modelling behaviour.

Towards a Translational Framework Between Lived Law and Formal Law

Track: Emerging Research

Authors: Sylvia Crow (Northumbria University), Christopher Simmonds (Northumbria University)

Abstract: This paper explores how generative AI is being used by self-represented litigants in UK employment disputes. Drawing on an exploratory qualitative analysis of 128 publicly available Reddit discussions, it finds that users consistently distinguish between AI as a source of legal authority, which is widely distrusted due to hallucinations and inaccuracies, and AI as a practical tool for organising evidence, drafting narratives and improving communication. Building on these findings, the paper introduces an analytical distinction between “lived law” and “formal law” and argues that generative AI should be understood as translational infrastructure operating between these domains rather than as an automated lawyer. The paper proposes that responsible legal AI should be evaluated not only for doctrinal accuracy but also for its ability to faithfully transform lived experiences into institutionally recognisable legal representations, offering a new perspective for research on access to justice and AI-assisted legal participation.

Fast and Robust Simulation-Based Inference with Optimization Monte Carlo

Track: World Class Research Track

Authors: Vasileios Gkolemis (Harokopio University of Athens ATHENA Research Center), Christos Diou (Harokopio University of Athens), Michael Gutmann (University of Edinburgh)

Presenter: Michael Gutmann (University of Edinburgh)

Abstract: Bayesian parameter inference for complex stochastic simulators is challenging due to intractable likelihood functions. Existing simulation-based inference methods often require large number of simulations and become costly to use in high-dimensional parameter spaces or in problems with partially uninformative outputs. We propose a new method for differentiable simulators that delivers accurate posterior inference with substantially reduced runtimes. Building on the Optimization Monte Carlo framework, our approach reformulates inference for stochastic simulators in terms of deterministic optimization problems. Gradient-based methods are then applied to efficiently navigate toward high-density posterior regions and avoid wasteful simulations in low-probability areas. A JAX-based implementation further enhances the performance through vectorization of key method components. Extensive experiments, including high-dimensional parameter spaces, uninformative outputs, multiple observations and multimodal posteriors show that our method consistently matches, and often exceeds, the accuracy of state-of-the-art approaches, while reducing the runtime by a substantial margin.

Exploring Human Behaviours of Stress through Virtual Reality Interactions using Accessible, Embedded Sensors and User-Personalised Machine Learning

Track: World Class Research Track

Authors: Saskia Davies (Swansea University), Tom Owen (Swansea University), Sean Walton (Swansea University)

Abstract: Virtual reality (VR) is increasingly applied in therapy due to its immersive and adaptable qualities, often augmented with sensors to facilitate the behavioural analysis of users. However, research in this domain commonly employs complex or intrusive setups that hinder real-world impact, arising from practicality limitations. In response, we aim to explore the capability of standalone VR sensors in stress detection. We gather a physiological dataset through VR user studies with 41 participants. A personalised random forest model is created for each user, trained against a 'stress index' produced by a medical-grade sensor. These models achieve 83.77% accuracy on average, and feature importance analysis identifies hand behaviours to be most significant for stress prediction. Further preliminary analysis suggests that high-velocity hand movements may be indicative of high stress levels. Results demonstrate how VR stress prediction can be achieved without complex setups, while informing the modalities and behaviour analysis of future systems.

Neural Mutual Information Estimation with Vector Copulas

Track: World Class Research Track

Authors: Yanzhi Chen (University of Cambridge), Zijing Ou (Imperial College London), Adrian Weller (University of Cambridge), Michael Gutmann (University of Edinburgh)

Presenter: Michael Gutmann (University of Edinburgh)

Abstract: Estimating mutual information (MI) is a fundamental task in data science and machine learning. Existing estimators mainly rely on either highly flexible models (e.g., neural networks), which require large amounts of data, or overly simplified models (e.g., Gaussian copula), which fail to capture complex distributions. Drawing upon recent vector copula theory, we propose a principled interpolation between these two extremes to achieve a better trade-off between complexity and capacity. Experiments on state-of-the-art synthetic benchmarks and real-world data with diverse modalities demonstrate the advantages of the proposed estimator.

Magnitude Distance: A Geometric Measure of Dataset Similarity

Track: World Class Research Track

Authors: Sahel Torkamani (School of Informatics, University of Edinburgh), Henry Gouk (School of Informatics, University of Edinburgh), Rik Sarkar (School of Informatics, University of Edinburgh)

Presenter: Sahel Torkamani (School of Informatics, University of Edinburgh)

Abstract: Distance metrics are fundamental tools shaping machine learning methods, from clustering and classification to generative modelling. We introduce a new method to measure the distance and similarity between two sets of data points. To achieve this, we use a mathematical tool called "Metric Magnitude," which can capture the geometric structure of data at different scales. Magnitude captures the structure of the data space through a tunable scaling parameter: a large scale focuses more on fine local details, while a small scale captures the global structure. The distance defined using magnitude can be scaled accordingly for a local or global view. Theoretically, we demonstrate fundamental properties of this distance, including its behavior in high-dimensional settings. While standard distances tend to collapse and lose their ability to distinguish differences in high dimensions, magnitude distance maintains discriminative power. Also, we show that under mild assumptions, our measure is robust to outliers: adding or removing a datapoint does not cause extreme changes. We also explain how the magnitude-distance relation behaves across different scaling regimes. As a proof of concept, we used the distance aggregated across different scales as a core guiding signal (the loss function) to train a new generative AI model called MagGN. By learning to generate new images with metric magnitude values and geometric properties similar to those of the training points, MagGN shows strong performance -- in some cases better than existing methods.

A Hybrid Heuristic-LLM System for Autonomous FPGA Timing Closure

Track: Emerging Research

Authors: Matthew Youngman (University of Edinburgh), Shu Gu (University of Edinburgh), Cristian Sestito (University of Edinburgh), Yuqin Zhao (University of Edinburgh), Ramesh Krishnamurthy (University of Edinburgh), Ali Siddique (University of Edinburgh), Danial Chitnis (University of Edinburgh), Themis Prodromakis (University of Edinburgh)

Abstract: Closing timing on Field Programmable Gate Array (FPGA) designs often requires solutions tailored to specific violation causes, yet vendor tools provide only generic flows, leaving diagnosis and remedy selection to expert engineers. We address this bottleneck with two contributions: a catalogue of optimisation flows, each targeting a distinct diagnosed cause of timing failure, and a system that automatically selects the required flow or sequence of flows for a given design. The system combines fast heuristics with LLM reasoning when signals are ambiguous; critically, all modifications are executed only through pre-built, validated flows, with the LLM restricted to selection, ensuring safety and reversibility. While demonstrated on timing, the framework could extend to objectives such as power and area.

Supporting “Stop and Watch” in Elderly Care with Socially Assistive Robots: Insights from a Participatory Design Workshop

Track: World Class Research Track

Authors: Tan Viet Tuyen Nguyen (University of Southampton), Athina Georgara (University of Southampton), Lokesh Singh (University of Southampton), Jayati Deshmukh (University of Southampton), Sean Davey (St Monica Trust), Paul N. Tisdale (St Monica Trust), Sarvapali D. Ramchurn (University of Southampton)

Abstract: “Stop and Watch” is an early-warning tool adopted by the UK NHS and is widely used in elderly care settings. The tool helps caregivers to recognise abnormal changes in residents’ health. Despite its clinical value, the process remains highly manual, workload-intensive, and vulnerable to missed observations, particularly in environments facing staff shortages, frequent staff rotations, and increasing care demands. We argue that AI-based systems such as Socially Assistive Robots (SARs) and Assistive Living Technologies (ALTs) offer promising avenues for supporting and enhancing the “Stop and Watch” tool. However, designing such systems requires a multidisciplinary effort to establish a comprehensive understanding of current practices and the priorities and concerns of all relevant stakeholders. This paper presents insights from a participatory design workshop held in a care home in the UK to explore how SARs and ALTs could meaningfully support the “Stop and Watch” tool, understand stakeholders’ expectations, perceived benefits, and concerns regarding deployment in this sensitive context.

GreenLine: A Delay-tolerant Mechanism Design for Grid Capacity Allocation

Track: World Class Research Track

Authors: Athina Georgara (University of Southampton), Jayati Deshmukh (University of Southampton), Sarvapali D. Ramchurn (University of Southampton)

Abstract: This work presents GreenLine, a new auction-based mechanism for integrating renewable power plants (RPP) in the power grid. We address a challenging task, where RPP owners want to maximise their profits by installing new power plants in the grid; while, the Distribution Network Operator (DNO) seeks to maximise the generated power while reducing potential installation delays due to growing power demand. This paper formulates this task as a multi-agent system, studies its useful properties such as incentive compatibility, individual rationality and economical efficiency, and discusses GreenLine under different deployment variations.

ProToMEx: Rapid, Interpretable Explanations via Structured Representations

Track: Emerging Research

Authors: Athina Georgara (University of Southampton), Adarsh Valoor (University of Southampton), Sarvapali D. Ramchurn (University of Southampton)

Abstract: Post-hoc explainers for machine-learning classifiers predominantly rely on feature attribution, assigning an importance score to each feature. Such scores rarely articulate the combinatorial patterns that drive a decision. We introduce ProToMEx, a model-agnostic explainer that leverages Probabilistic Topic Models (PTMs) to learn latent topics representing distinct, high-level reasons for a classification. Our preliminary empirical evaluation shows that ProToMEx matches the fidelity of SHAP and LIME while reducing the amortised cost of local explanations by roughly 30–40× on standard tabular and synthetic datasets.

Towards a Structured Framework for Large-Scale AI-Driven Thematic Analysis on Domain-Specific Customer Feedback

Track: World Class Research Track

Authors: Jade Logan (Swansea University), Nicholas Micallef (Swansea University), Daniele Doneddu (Swansea University), Kevin McLafferty (Swansea University), Muneeb Ahmad (Swansea University)

Presenter: Jade Logan (Swansea University)

Abstract: The recent surge in Large Language Models (LLMs) has led to their use in large-scale qualitative analysis, yet many approaches remain opaque, methodologically weak, and low in reflexivity, limiting their validity. This paper presents a structured framework for fully automated, LLM-supported thematic analysis grounded in Braun and Clarke's method, evaluating two prompting strategies: iterative analysis and progressive batch expansion. A pilot study with financial-services customer experience (CX) feedback compares LLM-generated themes against independently coded human themes across Net Promoter Score comments. Results show that iterative analysis yields Jaccard indices of 0.71—0.86 with human themes, compared to 0.44—0.71 for batch processing, with higher alignment for dominant CX issues but weaker coverage of lower-frequency concerns. The framework offers a structured, scalable approach to CX analysis and clarifies where LLM-led thematic analysis converges with—and diverges from—human judgement.

Risk-averse optimization of genetic circuits under uncertainty

Track: World Class Research Track

Authors: Michal Kobiela (University of Edinburgh), Diego A. Oyarzún (University of Edinburgh), Michael Gutmann (University of Edinburgh)

Abstract: Engineering biological systems with specified functions requires navigating an extensive design space, which is challenging to achieve with wet-lab experiments alone. To expedite the design process, mathematical modeling is typically employed to predict circuit function in silico ahead of implementation, which, when coupled with computational optimization, can be used to automatically identify promising designs. However, circuit models are inherently inaccurate, which can result in suboptimal or non-functional in vivo performance. To mitigate this, we propose combining Bayesian inference, Thompson sampling, and risk management to find optimal circuit designs. Our approach employs data from non-functional designs to estimate the distribution of model parameters and then employs risk-averse optimization to select design parameters that are expected to perform well, given parameter uncertainty and biomolecular noise. We illustrate the approach by designing adaptation circuits and genetic oscillators using real and simulated data, with models of varied complexity.

The Local--Global Alignment Paradox: Evaluating Distributed Safety Reconstruction in Multi-Agent Systems

Track: Emerging Research

Authors: Zhongjiang Yao (King's College London), Zhiqin Huang (Institute of Information Engineering, Chinese Academy of Sciences), Zuming Chang (Institute of Information Engineering, Chinese Academy of Sciences), Yepeng Yao (Institute of Information Engineering, Chinese Academy of Sciences), Guoyi Li (Institute of Information Engineering, Chinese Academy of Sciences), Yali Du (King's College London)

Abstract: This emerging research paper studies a safety blind spot in multi-agent LLM systems: harmful intent can be distributed across agents, roles, and turns so that each individual message appears benign, while the assembled role-labelled dialogue reconstructs a restricted goal. We introduce MASCSBench, a diagnostic benchmark of 10,162 multi-agent conversations across seven risk categories, and compare local, flat-context, and structure-aware safety monitors. Our results show that single-response scoring substantially under-detects distributed risks, while a dialogue-graph protocol improves detection by tracking how evidence accumulates across roles and time. The paper argues that agent safety should be evaluated at the level of the assembled workflow, not only at the level of isolated responses.

Machine learning-based predictions of healthcare contacts following emergency hospitalisation using electronic health records

Track: World Class Research Track

Authors: Konstantin Georgiev (Institute of Neuroscience and Cardiovascular Research, Queen's Medical Research Institute, University of Edinburgh), Dimitrios Doudesis (Institute of Neuroscience and Cardiovascular Research, Queen's Medical Research Institute, University of Edinburgh), Joanne McPeake (Department of Public Health and Primary Care, The Healthcare Improvement Studies Institute, University of Cambridge), Jacques Fleuriot (Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh), Susan Shenkin (Ageing and Health Research Group and Advanced Care Research Centre, Usher Institute, Edinburgh BioQuarter, University of Edinburgh), Nicholas Mills (Institute of Neuroscience and Cardiovascular Research, Queen's Medical Research Institute, University of Edinburgh), Atul Anand (Institute of Neuroscience and Cardiovascular Research, Queen's Medical Research Institute, University of Edinburgh)

Abstract: Emergency departments increasingly struggle to allocate appropriate care for older patients. Current triage in urgent hospitalisations fails to clearly distinguish between healthcare needs during the critical 72-hour post-admission window, risking delays in treatment commencement and subsequent poor patient outcomes. To address this issue, we examined the potential of Machine Learning algorithms to identify in-hospital healthcare contacts among older patients after emergency admission, using linked electronic health record data from three hospitals in South-East Scotland. We developed and validated XGBoost (gradient-boosted) models from 98,242 urgently hospitalised older patients, trained on frailty markers, nursing risk assessments (4AT delirium score, Waterlow pressure ulcer score, malnutrition), lab results, medication history, and prior diagnoses. Mean absolute percentage errors for nursing and rehabilitation contact prediction improved from 49% at ED arrival to 34% by 72 hours. Secondary classification models performed well on binary endpoints: AUCs of 0.89 for identifying patients needing urgent geriatric services, 0.83 for rehabilitation need, and 0.85 for predicting in-hospital death, assessed within the first 24 hours of admission. After the first day of admission, the pressure ulcer risk score was a prominent predictor of increased healthcare contacts. Leave-one-hospital-out validation showed that the model generalised reasonably across sites. Our study demonstrates that integrating routinely collected healthcare data with AI algorithms can facilitate faster allocation of care in urgently hospitalised older patients, simultaneously supporting more precise triage to specialist geriatric services.

If LLMs Misbehave, Will We Even Be Able to Tell?

Track: Emerging Research

Authors: Austin Long (University of York), Colin Paterson (University of York)

Abstract: Training large language models to perform latent reasoning may improve performance, but the lack of legible reasoning traces challenges model interpretability and safety assurance. This paper investigates whether existing interpretability methods can be used to effectively decode latent reasoning into a legible form. We find that the logit lens can substantially improve the legibility of COCONUT's latent reasoning tokens, but explicit reasoning is still significantly more legible.

Human Sign-Off as an Accountability Sink for AI-Generated Records in UK General Practice

Track: Emerging Research

Authors: Maureen Mwadime (Northumbria University)

Presenter: Maureen Mwadime (Northumbria University)

Abstract: Ambient voice technology (AVT) is being adopted across UK general practice on the strength of a simple settlement [4]: the AI drafts, the clinician reviews and signs, and accountability stays with the human whose name is on the note. This short paper asks whether that settlement holds. The sign-off may not retain human accountability so much as manufacture it, converting a system's authorship into a person's liability through a single act of approval made under time pressure, conditions under which a meaningful minority of GPs already report significant or critical errors in AI-generated notes [3]. The clinician holds the pen in name; the system holds it in substance. The argument developed here is that human oversight, as currently configured in the AVT rollout [13], can function less as a safeguard than as an accountability sink [2]: a mechanism that absorbs the appearance of responsibility while doing little to secure its reality. The consequence is borne by the patient, who is written about by a third party, governed by a plan that assumes a review that may not be happening, and given no means to see or contest the account that becomes their medical history. The point is not that oversight is worthless, but that its ritual has been mistaken for its function, and that this mistake is built into the governance frameworks themselves. The reframing offered to the human-AI collaboration community is this: if sign-off is the point at which AI authorship is converted into the record, then "keep a human in the loop" is less a solution than a description of where the problem has been situated.

Worlding otherwise: Virtual production as a participatory method for imagining safer digital futures

Track: World Class Research Track

Authors: Helen Kennedy (University of Nottingham), Sarah Martindale (University of Nottingham), Anna Maria Piskopani (University of Nottingham), Richard Ramchurn (University of Nottingham), Aislinn Bergin (University of Nottingham)

Presenter: Helen Kennedy (University of Nottingham)

Abstract: This paper presents AI-accelerated virtual production (VP) as a participatory research method for imagining safer and more inclusive digital futures. Informed by deep research related to the toxicity of online environments, the research team created a bespoke multi-user VR world in which participants could explore alternative forms of interaction, cooperation, identity expression and digital governance. The virtual world was developed through a combination of creative dramaturgy, virtual production workflows and generative AI tools. A dramaturg and creative technologist translated participant insights about online misogyny, privacy, comfort, identity and control into a speculative environment that combined familiar, reassuring spaces with more fantastical and futuristic elements. Technically, the environment was built in Unity using Unity 6's Multi-user VR Template, which provided the basis for voice and text communication, shared networked objects and a lobby system. Generative AI was then used to accelerate asset creation. Buildings, characters, beach huts, trees and other visual elements were generated through a local Stable Diffusion workflow running via ComfyUI, with prompts repeatedly refined until the images were aesthetically and ethically appropriate. These still images were then processed through Polycam's image-to-mesh generator and optimised for use on Meta Quest 2 headsets. The resulting world took the form of an island landscape combining a futuristic city, naturalistic spaces and domestic refuges, described as merging cyberpunk with a friendly Animal Crossing-style aesthetic. The paper treats AI and VP as situated research tools. The workflow enabled rapid prototyping, but also exposed the biases built into generative systems, particularly around racialised and sexualised avatar generation. The virtual world became both a research environment and a critical apparatus for testing how safer, more caring digital spaces might be imagined, embodied and collectively negotiated.

Electronic Circuit Optimisation with Graph and Sheaf Neural Networks

Track: Emerging Research

Authors: Alexandros D. Keros (The University of Edinburgh), Eva Maria Hita Sogorb (The University of Edinburgh), Cristian Sestito (The University of Edinburgh), Themis Prodromakis (The University of Edinburgh)

Presenter: Eva Maria Hita Sogorb (The University of Edinburgh)

Abstract: Moore's law refers to the observation that the num- ber of transistors in an integrated circuit doubles ev- ery two years [1]. Since 1965, the growth in integrated circuit complexity had accurately followed this trend, until it started to plateau due to economical and physi- cal barriers. Nowadays, the main physical constraints affecting system performance are related to power, heat generation, and the physics of complementary metal-oxide semiconductor (CMOS) transistor switch- ing, all of which act as limiting factors [2]. In terms of economic limitations, Gordon Moore also observed a second law that described how the exponential in- crease in fabrication costs would become a limiting factor before physics [3]. As Moore's law becomes less applicable to the current technological development situation, new paradigms are needed to ensure power, performance and area savings (PPA) [4] for maintaining modern circuit-scaling demands. Approximate computing arises as a possible answer to this challenge, exploit- ing the trade-off between computational accuracy, power, and area requirements [5]. Gate-level pruning (GLP) stems from this paradigm by omitting gates from a circuit design while striving to maintain task- specific accuracy requirements. Current examples of tools that perform GLP require a time-consuming iterative simulation process. For example, a previ- ous study [6] presented a probabilistic pruning-based tool [7] that is integrated in the digital circuit design flow to perform automatic GLP. This tool requires gate-level hardware simulations to be run in order to obtain a Switching Activity Interchange Format (SAIF) file to inform the decision of which wires are to be pruned. However, a more efficient approach that al- lows to skip the iterative simulation step is needed. Artificial intelligence has been successfully ap- plied across several domains to automate complex tasks by making accurate predictions, leading to sig- nificant time savings in fields such as healthcare [8]. Graph Neural Networks (GNN) provide the frame- work for performing learning and inference on com- binatorial data, with successes in antibiotic discov- ery [9], where graphs encode molecular networks and their interactions, and in analogue circuit repre- sentation [10]. Sheaf Neural Networks (SNNs) appear as generalizations of GNNs,

where information on vertices and edges is modeled as data in appropri- ate vector spaces, with consistency maps between them, forming what are known as cellular sheaves [11] [12]. SNNs have been successfully used in applica- tions such as recommendation systems, outperform- ing GNNs [13]. GNNs have been explored in EDA problems such as congestion prediction [14], testability analysis [15], and classification of sub-circuits in netlists for reverse engineering (RE) purposes. For example, DeepGate is a GNN-based solution for the representation learn- ing of logic gates [16]. The purpose of this model, as well as that of its subsequent versions such as Deep- Gate2 [17], is to create a general-purpose neural rep- resentation framework for logic circuits, as well as a circuit-aware GNN architecture that can be applied as the basis to solve several EDA problems. However, even the latest version of the algorithm, DeepGate4, is susceptible to the inherent limitations presented by GNNs, such as oversmoothing, which suggests that another approach might be necessary. We hereby show the representation of digital cir- cuits as sheaves (Figure 1), and propose the use of SNN to learn and automate GLP. By representing op- erational characteristics, power demands, and area requirements concisely as stalks, and circuitry opera- tions as restriction maps of a circuit sheaf, we propose to learn a sheaf diffusion operator that performs GLP.

Reframing Barriers to AI Adoption: An Emerging Perspective on Responsibility in High-Stakes Environments

Track: Emerging Research

Authors: Roger Kirby (University of Portsmouth), Mark Xu (University of Portsmouth), Zeeshan Bhatti (University of Portsmouth)

Presenter: Roger Kirby (University of Portsmouth)

Abstract: High-stakes environments are characterised by risk of harm, regulatory intensity and accountability requirements. Whilst Responsible Artificial Intelligence (RAI) is often presented as introducing distinct governance challenges, organisations in these domains have long operated under heightened responsibility requirements. This paper explores how barriers to AI adoption in high-stakes environments are characterised and proposes a three-layer analytical lens combining barrier groups, constraint type (AI, RAI and Hybrid) and the Technology–Organisation–Environment (TOE) framework. The approach provides a basis for investigating how responsibility-related requirements are expressed within AI adoption barriers and forms the analytical foundation of a PRISMA-guided Systematic Literature Review (SLR).

From Radar to Report: Auditing Claim-Level Faithfulness in LLM Tremor Reports

Track: Emerging Research

Authors: Haoxuan Li (Institute of Sensors, Signals and Systems, Heriot-Watt University), Wenda Li (Institute of Sensors, Signals and Systems, Heriot-Watt University)

Presenter: Haoxuan Li (Institute of Sensors, Signals and Systems, Heriot-Watt University)

Abstract: Contactless mmWave radar can measure hand tremor, and a language model can turn the per-hand measurements into a clinician-facing report. But it can say more than the measurements support, adding diagnoses, recommendations, or severities the radar never recorded, a safety concern for clinical language generation. Because the radar record is exact and shown to the model, faithfulness can be checked deterministically: a frozen rule-based extractor splits each report into typed atomic claims, and a deterministic closed-world oracle labels each one supported, unsupported, misleading, or harmless, with no retrieval and no learned judge. We audit eleven models on a pre-specified adversarial suite, and gpt-5.5 on a servo-phantom bench and TENS-induced surrogate tremor. Ungrounded prompting yields unsupported or misleading claims in 29 to 60% of claims and 83 to 100% of reports, mainly unmeasured recommendations and invented diagnoses. Grounding removes them for five of the eleven models and shrinks them for three more, but three resist it, and one worsens under it; coverage of the measured facts rises from 86 to 95%, against 100% for the template. A numeric string-matching baseline misses every interpretive error the typed oracle catches.

Tempered Guided Diffusion

Track: Emerging Research

Authors: Andreas Makris (Lancaster University), Paul Fearnhead (Lancaster University), Chris Nemeth (Lancaster University)

Abstract: Training-free conditional diffusion methods adapt a pretrained unconditional diffusion model to new inverse problems at sampling time. This flexibility is attractive when paired training data or task-specific conditional models are unavailable, but current multi-trajectory heuristics often allocate computation inefficiently: poor trajectories may be propagated to completion, while additional function evaluations along a single trajectory may not recover from early mistakes. We propose Tempered Guided Diffusion (TGD), a sequential Monte Carlo framework that defines an annealed sequence of clean-signal posterior targets and uses diffusion states only as auxiliary variables for proposing and propagating particles. Particles are reweighted by incremental likelihood ratios, resampled, and propagated across noise levels so that later computation is concentrated on candidates plausible under both the generative prior and the observation. For expensive image inverse problems, Accelerated TGD prunes to a high-likelihood trajectory partway through sampling. Experiments show improved posterior approximation and favourable wall-clock speed--quality trade-offs over independent guided-trajectory baselines.

Fairness Explanations Improve Fairness in Human-AI Decision Making: An Empirical Evaluation of Parole Decision-Making

Track: Emerging Research

Authors: Matt Roach (Swansea University), Peter Daish (Swansea University), Alan Dix (Swansea University), Muneeb Ahmad (Swansea University)

Presenter: Peter Daish (Swansea University)

Abstract: Data-driven Artificial Intelligence (AI) systems are increasingly deployed in high-stakes public domains, yet they often reproduce harmful systemic so cial biases. While Explainable AI (XAI) has been widely adopted to improve the transparency of AI utility, little is known about how exposing an AI system's social bias to human decision-makers impacts hybrid human–AI outcomes. In this emerging research, we investigate the impact of visual "Fairness Explanations" on fairness, performance, and usability in human-AI decision-making. We contribute: (1) a design prototype and algorithmic methodology for generating visual fairness explanations; (2) a methodologi cal approach for constructing a socially bias-free dataset from a historically biased source (COMPAS); and (3) empirical evidence from a counterbal anced within-subjects study (n = 48) in a parole decision-making context. Participants were evaluated across three conditions: unassisted, AI-assisted, and Fairness-Explained AI-assisted (Fair-XAI). Our findings demonstrate that fairness explanations significantly reduce multiple dimensions of bias (sex and ethnicity) compared to unassisted human decision-making, while also improving the accuracy of positive predictions. However, these fairness gains currently come at a cost to system usability.

Radicalisation Amplified by LLM Sycophancy

Track: Emerging Research

Authors: Hirokazu Hatta (University of Bath), Harish Madabushi (University of Bath), Laura Smith (University of Bath)

Abstract: Large language models (LLMs) increasingly serve as conversational partners, yet their sycophantic tendency may have unintended consequences. We tested whether interacting with an LLM causes radicalisation while discussing the immigration in Britain. UK nationals (N = 420) completed an identity-salience writing task (Social, Self, or Control) before discussing immigration with a non-partisan LLM. Participants' political intentions declined across all conditions, suggesting that a neutral chatbot tempered their readiness to act. Prejudice toward immigrants showed no overall change, but emotional validation moderated the anger–prejudice link: expressed anger predicted rising prejudice only when validation was high. These findings suggest that validation did not generate extremity but rather converted existing anger into concerning attitudes, and the features that make LLMs engaging are not always beneficial.

Street Scenes: Public Appliances for GenAI Video in Informal Settlements

Track: World Class Research Track

Authors: Gavin Bailey (Swansea University), Dani Raju (Studio Hasi), Jennifer Pearson (Swansea University), Simon Robinson (Swansea University), Matt Jones (Swansea University)

Abstract: Generative AI is rapidly diffusing worldwide, yet access remains uneven. In informal settlements, barriers of cost, literacy and connectivity can exclude residents from AI-enabled self-expression. This paper presents Street Scenes: a public appliance for walk-up interaction with generative AI video in Dharavi, Mumbai. Inspired by the “Hole in the Wall” computers and previous Dharavi speech deployments, the system lets passers-by capture phone images, add voice-, button- and dial-based prompts, and generate short videos to view and leave locally. We report on ideation workshops, two Wizard-of-Oz prototypes and a 13-day in-situ deployment across Dharavi street locations. Findings show residents appropriating AI for play, self-presentation, small business promotion and community messaging, while also raising concerns about privacy, trust and misuse. We contribute: (1) a model for public AI appliances; (2) empirical insights into community engagement with generative AI; and, (3) design lessons for accessible, equitable and community-governed AI systems.

Less Redraw, More Explore: Suggestion and Completion for Sketch-to-Image

Track: World Class Research Track

Authors: Zeyu Zhao (Swansea University), Connor Rees (Swansea University), Gavin Bailey (Swansea University), Matt Jones (Swansea University), Simon Robinson (Swansea University), Jennifer Pearson (Swansea University)

Abstract: Sketch-to-image systems let users transform simple line drawings into realistic images, but current workflows force users into tedious redraw-regenerate cycles that slow creative exploration. We introduce two complementary interaction techniques that reduce iteration friction: AutoSketch, which extends partial sketches through AI-driven completions (pre-generation support), and BackSketch, which transforms generated images back into editable sketches at multiple abstraction levels (post-generation support). In a study with 30 participants, the results indicate that both techniques can improve exploration and expressiveness compared to a baseline sketch-to-image system, while AutoSketch also can increase users' sense of agency and co-creation with the AI. We contribute new evidence that shifting support before or after generation opens distinct pathways for balancing user control and system initiative. Together, our results establish pre- and post-generation assistance as a design space for co-creative sketch-to-image systems.

The ORBIT India Dataset: Understanding the Challenges of Collecting a Disability-First AI Dataset in Low-Resource Environments

Track: World Class Research Track

Authors: Gseu India (Swansea University), Martin Grayson (Microsoft Research), Cecily Morrison (Microsoft Research EMEA), Daniela Massiceti (Microsoft Research), Simon Robinson (Swansea University), Jennifer Pearson (Swansea University), Matt Jones (Swansea University)

Presenter: Jennifer Pearson (Swansea University)

Abstract: Computer vision systems are increasingly used by blind individuals to navigate their lives, helping, for example, locate objects such as doors or chairs. Yet these recognition systems do not work for many personal objects a blind user might want to find, such as keys or a special notebook. In response, efforts created personalized recognition systems, where individuals train their phones to identify and locate things, like a coffee mug or white cane, using example images/videos. However, these tools are trained on data from high-resource contexts, not necessarily reflecting India's material culture. This paper discusses the contribution of the ORBIT-India dataset, which extends these tools to the Indian context, home of the world's largest blind population. The ORBIT-India dataset comprises 105,243 images from 587 videos, representing 76 unique objects. We use this experience to examine dataset collection practices translated from high- to low-resource settings, providing recommendations to support cross-geography dataset collection.

PhyTENG: A physics-informed machine learning framework for modelling contact - separation triboelectric nanogenerators

Track: World Class Research Track

Authors: Shashank Mishra (University of Edinburgh), Ram Ramamoorthy (University of Edinburgh), Jacqueline M. Cole (University of Cambridge), Themis Prodromakis (University of Edinburgh)

Presenter: Shashank Mishra (University of Edinburgh)

Abstract: Triboelectric nanogenerators (TENGs) are promising for harvesting mechanical energy and enabling self-powered sensing. Their electrical performance depends on complex interactions among materials, device architecture, and operating conditions and it is challenging to make accurate prediction models. In this work, we present PhyTENG, a physics-informed machine learning (ML) framework to predict output voltage and current in contact-separation (CS) TENGs. A diverse experimental dataset was curated from literature and augmented with open-source materials databases and ML-based property predictors to obtain atomic- and surface-level physics-based descriptors, including electron affinity, ionisation energy, and triboelectric charge density. Four interpretable gradient-boosting models were then trained to relate these properties and other descriptors to their electrical performance. We further demonstrate the advantages of using physics-inspired descriptors relative to conventional categorical materials featurisation. Overall, the presented PhyTENG framework provides a generalisable predictive tool for TENG optimisation and a pathway to integrate AI-driven materials discovery with AI-driven device design.

A Diffusion Prior for Active Inference Planning

Track: World Class Research Track

Authors: Catherine Higham (University of Glasgow), Roderick Murray-Smith (University of Glasgow)

Presenter: Catherine Higham (University of Glasgow)

Abstract: This paper presents a method for combining Denoising Diffusion Probabilistic Models (DDPM) with Active Inference (AIF). The key idea is to leverage DDPM as a behavioural prior for inference within the AIF framework. This work is a novel application of DDPM to create a prior over actions, observations, and states for efficient planning and decision making. The DDPM is trained on behavourial observations in a training environment. Once trained, the DDPM can be conditioned on observations from a new environment without the need for retraining. The agent is then equipped to imagine likely future scenarios and focus on regions of the space which are most relevant. We show that estimates of the probability of future states can be used to provide inexpensive estimates of the Expected Free Energy. In addition, as the dimensions of the state space increase, the computational cost of generating policies is offset by a reduction in the number of policies to be considered. Together, these advantages overcome a major computational bottleneck. In summary, behavioural observations are efficiently encoded as data to train a DDPM, providing a simple but powerful prior for behaviour in a new environment.

A Minimal Model of Bounded Trade-Off Screening in Multi-Attribute Choice

Track: Emerging Research

Authors: Manisha Dubey (University of Edinburgh), Anirban Sarkar (Cold Spring Harbor Laboratory, USA), Ram Ramamoorthy (University of Edinburgh)

Presenter: Manisha Dubey (University of Edinburgh)

Abstract: Human decision-making often involves choosing between multi-attribute alternatives, yet classical models assume fully compensatory utility aggregation despite evidence that people reject options with poor performance on critical attributes. We propose a bounded trade-off reasoning framework in which decisions are governed by a screening process that evaluates the balance between gains and losses across attributes. The model introduces a trade-off tolerance parameter that controls acceptable imbalance and can vary across contexts. Through simulation, we show that this mechanism produces preference patterns that differ from standard utility-based models and captures context-dependent variation in trade-off behavior. These results establish bounded trade-off screening as a plausible computational mechanism for multi-attribute choice and generate testable predictions for future behavioral studies.

Computational Agents for Air Traffic Control: Explainability, Interpretability and Trust

Track: Emerging Research

Authors: Robert Procter (University of Warwick), Mark Rouncefield (University of Siegen), Richard Everson (Exeter University), George De Ath (Exeter University), Nick Pepper (Alan Turing Institute), Edward Henderson (Alan Turing Institute), Elhassan Mohamed (Alan Turing Institute), Eseoghene Benjamin (Alan Turing Institute), Ben Carvell (NATS)

Presenter: Robert Procter (University of Warwick)

Abstract: We report findings from an ongoing investigation into transparency techniques for computational Air Traffic Control agents. Air Traffic Controller Officers (ATCOs) observed two types of agent in simulated air traffic controlling exercises. ATCOs found the chosen transparency techniques helpful to some extent: revealing how agents were making controlling decisions was an essential foundation for trust. However, the ATCOs' trust depended on whether they could be made sense of in a real-time, operational context and the agents' actions where judged to be professionally competent.

Designing a Community-Governed Tool to Mediate AI Use in Minority Ethnic Communities: A Participatory Approach

Track: Emerging Research

Authors: Nicholas Micallef (Swansea University), Yan Wu (Swansea University)

Presenter: Nicholas Micallef (Swansea University)

Abstract: AI systems increasingly mediate how minority ethnic communities operate online, yet those communities are rarely given control over how such systems behave on their behalf. We present an emerging, work-in-progress account of a participatory approach to design a community-governed AI tool that acts as decision-support for civic organisations rather than as an automated agent. Working with two minority ethnic community organisations in Wales, we treat the design problem not as "building a better classifier" but as distributing governance to the people the system serves. We describe a participatory method that combines AI literacy with norm elicitation in a single workshop, so that community members can reason about both what they need and the technical levers available to deliver it; and we describe how the elicited norms are translated into inspectable, editable system components rather than left as advisory input. Our key provocation is that for AI to collaborate well with minority ethnic communities, governance capacity-the demonstrated ability of community members to inspect, contest, and correct the system-should be treated as a primary design outcome, not a side effect. We propose an evaluation that would test this directly in a controlled lab study with our partner communities.

Designing Inclusive and Sustainable Multilingual AI through Small Language Models

Track: Emerging Research

Authors: Anna Pramod (University of Leeds), Vivin Rajasekharan Nair (UST)

Presenter: Anna Pramod (University of Leeds)

Abstract: Artificial Intelligence (AI) systems increasingly shape communication, education, digital participation, and access to information. However, many multilingual AI systems remain heavily dependent on large-scale computational infrastructure and disproportionately support high-resource languages, creating barriers for low-resource linguistic communities and constrained deployment environments. These limitations affect not only access to AI technologies, but also equitable and sustainable participation in modern computing systems. This work explores whether Small Language Models (SLMs), combined with parameter-efficient optimisation strategies, can support more inclusive and sustainable multilingual AI. A pilot study was conducted by fine-tuning TinyLlama for multilingual Indic sentiment analysis, incorporating Low-Rank Adaptation (LoRA) and deployment-oriented optimisation techniques. Experimental analysis compared the proposed approach against multilingual BERT baselines while evaluating predictive performance, memory usage, and deployment feasibility. Results demonstrate that optimised SLMs can achieve competitive multilingual performance with significantly reduced computational requirements. This work positions multilingual AI not only as a Natural Language Processing (NLP) challenge, but also as a human-centred computing problem involving accessibility, computational equity, and sustainable participation in digital systems.

Emerging Research in the Evaluation of Large Language Models in Law

Track: Emerging Research

Authors: Maria Liakata (Queen Mary University London), Robert Procter (University of Warwick), Xingwie Tan (University of Warwick), Aislinn Bergin (University of Nottingham), Jiahong Chen (University of Sheffield), Maria Waheed (University of Warwick), Tom Sorell (University of Warwick), Jenny Chim (Queen Mary University London), Serene Chi (University of Warwick), Joshua Kelsall (University of Warwick)

Abstract: Large language models (LLMs) are increasingly proposed and adopted as tools to address pressing challenges in the legal sector, including case backlogs and unequal access to legal services. While standardised technical metrics and evaluation techniques have been used to evaluate LLMs in legal use cases, socio-technical evaluation of LLMs remains largely underexplored. Drawing on a rapid evidence review of 140 studies on LLM evaluation in legal use cases (Kelsall et al. 2026), alongside our project's (AdSoLve) ongoing engagement with key stakeholders in the legal domain to understand legal AI adoption in practice, this paperidentifies a set of emerging themes concerning central gaps in the socio-technical evaluation of LLMs in the legal domain, which inform the AdSoLve project aim of providing robust evaluation metrics for LLMs in law.

Teachers, AI, and Professional Judgement: Exploring How Teachers Make Sense of AI Use and Non-Use in Practice

Track: Emerging Research

Authors: Joella Lynch (Northumbria University)

Presenter: Joella Lynch (Northumbria University)

Abstract: Whether, when, and how to use AI are now questions many teachers are asking in their daily practice. Yet most research into AI in education has focused on the technology itself not on how teachers are making sense of these decisions. This study explores how teachers are exercising professional judgement in relation to using or not using AI in their practice, through reflective workshops with 88 teachers across primary, secondary, further education, career-related learning, and international settings. Emerging analysis suggests that teachers are not simply adopting or rejecting AI, but are negotiating tensions between innovation, professional values, and what it means to be a teacher.

Taming the Sigmoid Bottleneck: Provably Argmaxable Sparse Multi-Label Classification

Track: World Class Research Track

Authors: Andreas Grivas (The University of Edinburgh), Antonio Vergari (The University of Edinburgh), Adam Lopez (The University of Edinburgh)

Presenter: Andreas Grivas (The University of Edinburgh)

Abstract: Sigmoid output layers are widely used in multi-label classification (MLC) tasks, in which multiple labels can be assigned to any input. In many practical MLC tasks, the number of possible labels is in the thousands, often exceeding the number of input features and resulting in a low-rank output layer. In multi-class classification, it is known that such a low-rank output layer is a bottleneck that can result in unargmaxable classes: classes which cannot be predicted for any input. In this paper, we show that for MLC tasks, the analogous sigmoid bottleneck results in exponentially many unargmaxable label combinations. We explain how to detect these unargmaxable outputs and demonstrate their presence in three widely used MLC datasets. We then show that they can be prevented in practice by introducing a Discrete Fourier Transform (DFT) output layer, which guarantees that all sparse label combinations with up to k active labels are argmaxable. Our DFT layer trains faster and is more parameter efficient, matching the F1@k score of a sigmoid layer while using up to 50% fewer trainable parameters. Our code is publicly available at https://github.com/andreasgrv/sigmoid-bottleneck.

A Generalizable Inverse-Design Framework for Multi-Port Microwave Circuits

Track: Emerging Research

Authors: Lu Qian (Heriot-Watt University), Haijun Fan (Heriot-Watt University)

Presenter: Lu Qian (Heriot-Watt University)

Abstract: Multi-port microwave circuits are essential components in wireless front ends, but their direct synthesis remains difficult because of the limited applicability of analytical design methods and the highly nonlinear relationship between circuit geometry and electromagnetic response. This work presents a generalizable deep-learning inverse-design framework that combines data-driven EM modelling with physics-based network reduction. A convolutional neural network (CNN) predicts the frequency-dependent multi-port scattering matrix of a pixelized circuit, while a binary genetic algorithm (GA) performs geometry optimisation using the CNN as a fast EM emulator. The predicted full-port response is transformed into task-specific reduced-port responses through a network-reduction operator supporting different port selections and terminations. As a result, a single dataset and trained CNN can be reused for multiple circuit functions without additional EM data generation or retraining.

Kernel-based Optical Computing in High-dimensional Spaces

Track: Emerging Research

Authors: Man Fung Ho (Aston University), Egor Manuylovich (Aston University), Sergei K Turitsyn (Aston University), David Saad (Aston University)

Presenter: Man Fung Ho (Aston University)

Abstract: We propose a new approach in optical computing based on mapping data into a high-dimensional feature space. The approach combines task-specific nonlinear transformation with vector–matrix multiplication, both implementable using optical devices. Performance is demonstrated on regression and classification benchmarks, showing strong results.

SHAF: Small Language Model Integrated with Motion Modality for Multimodal Interaction

Track: World Class Research Track

Authors: Aamir Ahmad Ansari (University of Southampton), Tan Viet Tuyen Nguyen (University of Southampton), Sarvapali D. Ramchurn (University of Southampton)

Presenter: Tan Viet Tuyen Nguyen (University of Southampton)

Abstract: Multimodal interaction plays a vital role in human–AI interaction, enabling robots or AI agents to interpret human input from multiple sensory channels and respond through diverse communication modalities. This paper introduces SHAF, an LLM-based multimodal model capable of handling text, image, and human motion as both input and output modalities across different multi-turn conversational settings. In SHAF, vector quantization is employed to convert images and human motion into an aligned set of tokens, followed by pre-training and instruction fine-tuning of a small Large Language Model (LLM) on our newly created SHAF dataset. Experimental results demonstrate that SHAF achieves competitive performance in text-to-motion and motion-to-text tasks in comparison to relevant works, while handling an additional modality and supporting a broader range of tasks. This research contributes an LLM-based multimodal approach, with the aim of fostering deeper exploration of human motion modality in LLMs within the context of HRI and related domains.

BioCOMPASS: Integrating Biomarkers into Transformer-Based Immunotherapy Response Prediction

Track: World Class Research Track

Authors: Sayed Hashim (University of York), Frank Soboczenski (University of York), Paul Cairns (University of York)

Presenter: Sayed Hashim (University of York)

Abstract: Datasets used in immunotherapy response prediction are typically small in size, as well as diverse in cancer type, drug administered, and sequencer used. Models often drop in performance when tested on patient cohorts that are not included in the training process. Recent work has shown that transformer-based models along with self-supervised learning show better generalisation performance than threshold-based biomarkers, but is still suboptimal. We present BioCOMPASS, an extension of a transformer-based model called COMPASS, that integrates biomarkers and treatment information to further improve its generalisability. Instead of feeding biomarker data as input, we built loss components to align them with the model's intermediate representations. We found that components such as treatment gating and pathway consistency loss improved generalisability when evaluated with Leave-one-cohort-out, Leave-one-cancer-type-out and Leave-one-treatment-out strategies. Results show that building components that exploit biomarker and treatment information can help in generalisability of immunotherapy response prediction. Careful curation of additional components that leverage complementary clinical information and domain knowledge represents a promising direction for future research.

Multimodal Embodied Support for International Students: Context-Aware Dialogue with a Virtual Avatar

Track: Emerging Research

Authors: Yuying Zhang (University of Southampton), Indu Bodala (University of Southampton), Sarvapali D. Ramchurn (University of Southampton), Tan Viet Tuyen Nguyen (University of Southampton)

Presenter: Tan Viet Tuyen Nguyen (University of Southampton)

Abstract: International students face language and cultural adaptation challenges, while campus support information is distributed across platforms. This paper presents a context-aware embodied conversational framework combining lightweight multimodal context awareness, retrieval-augmented generation, and adaptive dialogue within a virtual avatar interface. Visual cues support engagement awareness, while language input guides adaptive interaction across three functional modes: companionship support, spoken-English practice, and campus activity recommendation. Preliminary results show that retrieval augmentation improves lexical overlap and semantic grounding compared with non-retrieval baselines, with GPT-4o and GPT-3.5 achieving the strongest performance.

Prior-Supervised Instrument-Aware Rollouts for Surgical Robot World Models

Track: Emerging Research

Authors: Yuhui Wan (School of Informatics, University of Edinburgh), Ram Ramamoorthy (School of Informatics, University of Edinburgh)

Abstract: Predictive models for robotic surgery must anticipate both future endoscopic images and how robot kinematics evolve under candidate controls. Existing surgical world models mainly forecast video or infer latent actions; they rarely test how the same observed context evolves under different recorded robot controls. We present an action-conditioned surgical robot world model. It jointly represents endoscopic observations, robot kinematics, and action-dependent transitions. Past frames and kinematics define the current state, while an explicit candidate action sequence drives the rollout of future images and robot states. Direct supervision of prior rollouts and reconstruction losses focused on instruments train the pathway used for open-loop prediction. Preliminary results on robotic cholecystectomy demonstrations show better perceptual fidelity and lower kinematic error than a latent state-space baseline. Predictions also change more strongly when actions are perturbed, indicating action responsiveness without establishing counterfactual accuracy. By explicitly conditioning rollouts on robot actions, the model supports action-dependent prediction rather than passive video forecasting. This makes the approach relevant to multimodal surgical AI, healthcare robotics, and model-based decision support.

Cripping the University Experience: AI-Assisted Making and Routes of Change with Students Living with Energy-Limiting Conditions

Track: Emerging Research

Authors: Christopher Jackson (Northumbria University)

Presenter: Christopher Jackson (Northumbria University)

Abstract: Students living with energy limiting conditions (ELCs) such as ME/CFS and Long COVID face barriers across the whole university experience, yet remain largely absent from HCI and accessibility research. Making and AI-assisted creation offer a route to agency, allowing students to create meaningful routes of change for themselves or the ELC student community. The dominant format for such making, the hackathon, is built on intensity, co-presence and time-bounding that make it fundamentally incompatible with crip time and fluctuating energy. This paper argues for \textit{cripping the hackathon}: a critical deconstruction of the format, with its reimagining left to co-design with the students it must serve. AI-assisted creation is positioned as the engine of low-energy making but is critically viewed for its risks to agency, critical thinking and ownership.

Machine-leaning assisted RHEED analysis: case studies applied to layer-by-layer thin-film growth

Track: Emerging Research

Authors: Yoshiko Nanao (University of St Andrews)

Presenter: Yoshiko Nanao (University of St Andrews)

Abstract: Two case studies demonstrating the universal utility and challenges of machine learning (ML) application to analysis of reflection high-energy electron diffraction images (RHEED). The complexicity of RHEED data provide rich insights on the crystal growth, while itself becomes a barrier to conduct detailed analysis. Here, we have applied both supervised/unsupervised-learning approaches to analyse RHEED data obtained from film growth of the naturally-layered compound $CuCrO_2$, with a particular focus on its potential to elucidate termination-specific features. Out proof-of-principle results already show that these methods hold the potential to establish a methodology for faster and more efficient ML-boosted optimisation process of thin-film materials growth.

Automated classification of natural habitats using ground-level imagery

Track: World Class Research Track

Authors: Mahdis Tourian (University of Exeter)

Abstract: This paper presents an automated approach for classifying natural habitats from ground-level imagery using deep learning. Accurate habitat classification is essential for biodiversity conservation, ecological monitoring, and land use planning, but conventional field based classification can be time-consuming and difficult to scale. While satellite imagery is widely used for habitat mapping, it still requires ground-level validation by trained ecologists. This study explores whether habitat types can be classified directly from ground-level photographs. Using 41,499 ecologist-captured RGB images from Natural England's Living England programme, the study develops a classifier for 16 broad habitat classes. A DeepLabV3-ResNet101 model was adapted from semantic segmentation to image-level classification and trained using transfer learning, image preprocessing, augmentation, class balancing, and five fold cross validation. The model achieved a mean F1-score of 0.63, with particularly strong performance for visually distinctive habitats such as Bare Sand, Coniferous Woodland, Built up Areas and Gardens, and Water. Performance was lower for visually mixed or ecologically ambiguous classes, including Unimproved Grassland and Fen, Marsh and Swamp, highlighting the limitations of imagery-only classification for habitats defined by subtle floristic or environmental features. The findings show that ground-level imagery can support scalable, repeatable habitat classification and may complement existing satellite based mapping and expert ecological assessment. The work also provides a prototype web application for classifying uploaded habitat images, supporting future applications in ecological monitoring, field validation, and citizen science.

Towards Human-Centred AI for Assistive Cooking for People Living with Dementia

Track: Emerging Research

Authors: Rimvydas Rubavicius (The University of Edinburgh), Nicole Meng-Schneider (The University of Edinburgh), Farhad Aghayev (The University of Edinburgh), Ram Ramamoorthy (The University of Edinburgh)

Presenter: Rimvydas Rubavicius (The University of Edinburgh)

Abstract: Using AI for the next-generation assistive technologies in social care presents challenges and opportunities: AI has the potential to augment users' skills in decline and preserve their autonomy, yet many attempts in deploying such technologies have faced struggles in long-term adoption by the users. A key facet of this problem is dyadic human-AI interaction in which the AI system is expected to adapt and provide personalised and contextualised assistance, based on the state of the environment as well as the user's capabilities and preferences. Achieving this requires human behaviour modelling and an AI system being a thoughtful partner with adaptive natural interaction. In this pursuit, we describe a prototype of an AI assistant for cooking for people living with dementia (PLWD) who benefit from cognitive scaffolding in the form of prompts. In the system prototype design process, we build on existing work in designing assistive technologies for PLWD for collaborative cooking, but extend the system by providing valuable features for both the PLWDs and the caregivers, yet at the same time empowering them to choose the degree of collaboration they wish to engage in.

Towards Learning-Assisted Optimisation of Electric Airport Ground Support Vehicle Operations

Track: Emerging Research

Authors: Yinghao Qin (Queen Mary University of London), John Korna (NATS), Marc Thomas (NATS), Richard Cannon (NATS), Jun Chen (Queen Mary University of London)

Abstract: Airports worldwide are increasingly electrifying their ground support vehicle fleets to reduce carbon emissions and improve operational sustainability. Examples include electric baggage tractors, aircraft towing vehicles, catering trucks, maintenance vehicles, and apron buses. While electric vehicles have been extensively studied in road transportation settings, the operational challenges associated with electric airport ground support fleets remain relatively underexplored. These fleets must operate within complex airside environments, satisfy time-critical service requirements, and manage limited charging resources. Existing approaches typically address routing, scheduling, and charging decisions separately, limiting their ability to support coordinated decision-making in increasingly electrified airport ecosystems. This paper outlines a research agenda for learning-assisted optimisation of electric airport ground support vehicle operations. The proposed framework integrates optimisation, machine learning, and digital-twin technologies to support planning and real-time operational decision-making. Three research directions are identified: integrated routing and charging optimisation, learning-assisted optimisation, and adaptive digital-twin-enabled operations. The proposed agenda aims to contribute to sustainable airport operations while advancing the integration of artificial intelligence and optimisation in transportation systems.

Towards Valid Generation of Financial Tabular Data with Diffusion Models

Track: Emerging Research

Authors: Mingxuan Yi (University of Bristol), John Cartlidge (University of Bristol)

Presenter: Mingxuan Yi (University of Bristol)

Abstract: Diffusion models have quickly become the best methods for generating synthetic tabular data, and recent work has started to use them in finance, where real data is hard to share and synthetic data is therefore valuable. We argue that the way these models are evaluated does not fit the financial setting. Current measures check statistical resemblance, such as how well each column and each pair of columns is reproduced, and how private the data is. They do not check whether a generated record follows the rules of the domain, for example that debits equal credits, that amounts are not negative, and that dates are in the right order. Recent results show that standard models often break such rules, and because later models are trained on this synthetic data, the errors can spread without being noticed. This work reviews diffusion models for tabular and financial data and explains why resemblance is not the same as validity. We then suggest two directions: an evaluation that reports the share of valid records, which we call the admissibility rate, and diffusion models that build domain rules into the generation process. Finally, we describe a planned study on public financial data, where we expect resemblance to stay high while the admissibility rate is much lower.

Voices of the Wear - Exploring AI's Role in Intangible Heritage Through Research-through-Design in North East England

Track: Emerging Research

Authors: Chris Corner (Northumbria University)

Abstract: Intangible cultural heritage in post-industrial regions, like in many places, is under threat, leaving younger generations at risk of not understanding how the social customs, triumphs, and knowledge of the past shapes their regional identity. Working in partnership with Culture House Sunderland, this PhD adopts a Research-through-Design methodology, with human-centred co-design principles to develop engaging AI systems for Sunderland's shipbuilding heritage in three case studies. By incorporating existing museological engagement theory into the design, the project aims to create a theoretical framework for AI engagement with intangible cultural heritage. Early co-design workshops and semi-structured interviews have highlighted six key themes for engagement, suggesting that relatable human stories and socially comfortable AI interactions are factors which allow for meaningful engagement with AI and intangible heritage.

Big Data's Blind Spots: Intersectionality and public opinion data on AI

Track: Emerging Research

Authors: Roshni Modhvadia (Ada Lovelace Institute)

Presenter: Roshni Modhvadia (Ada Lovelace Institute)

Abstract: As artificial intelligence becomes increasingly embedded in policy discourse, large-scale public attitudes data are often treated as authoritative evidence of what people want, fear, experience and expect from these technologies. Yet the same structural limitations identified in data science—data gaps, insufficient representation, reductive categories and the silencing of marginalised groups—also shape the production and interpretation of quantitative public opinion research on AI. This paper argues that, without examining the evidence production ecosystem as a whole, public attitudes data risks reproducing the inequalities they are often used to research. Drawing on practice-based experience of delivering nationally representative AI attitudes research, we consider tensions between policy demand for clear, representative statistics and the need to surface the views of people most likely to be affected by AI systems. Using survey data from the UKRI-funded (RAI and BRAID) survey How do people feel about AI? and an exploratory case study on facial recognition in policing and surveillance, we explore how combining innovations in intersectional analysis in other domains with critical frameworks for quantitative approaches can address some of the inequalities the methodological approach can reproduce.

Machine Learning EEG Biomarkers in SYNGAP1 Rodent Models

Track: Emerging Research

Authors: Nikoo Moradi (University of Edinburgh), Melissa Fasol (University of Edinburgh), Hollan Haule (University of Edinburgh), Javier Escudero (University of Edinburgh), Alfredo Gonzalez-Sulser (University of Edinburgh)

Presenter: Nikoo Moradi (University of Edinburgh)

Abstract: SYNGAP1-related disorder (SRD) is a neurodevelopmental disorder caused by mutations in the SYNGAP1 gene. SRD patients experience seizures and developmental problems. Novel treatments for SRD are in development, but reliable biomarkers for tracking their clinical efficacy have yet to be identified. We construct an XGBoost EEG classification pipeline in a SYNGAP1 rat model (9 SYNGAP1, 9 wild-type). Over 1000 spectral, complexity, and functional connectivity features were extracted. Aggregating epoch-level features to the animal-level achieved global area-under-ROC-curve of 0.889 and accuracy of 0.722 ($p = <0.0001$, 1000-permutation test). Four stable features, alpha/delta ratio, beta/theta ratio, and theta power, were selected in ≥3/18 folds as primary genotype discriminators. The identification of non invasive biomarkers of SRD and other neurodevelopmental disorders could be invaluable to track disease progression and determine treatment efficacy of novel therapeutics.

A Multi-Agent Framework for Extraction of Materials, Methodology and Performance data from Triboelectric Nanogenerator Literature

Track: Emerging Research

Authors: Sean Yalcin (University of Edinburgh), Shashank Mishra (University of Edinburgh), Chandrabhan Kushwah (University of Edinburgh), Themis Prodromakis (University of Edinburgh)

Abstract: We propose an evidence-grounded multi-agent pipeline for extracting materials, methods, and performance data from triboelectric nanogenerator literature. Specialised agents produce structured claims linked to source evidence, which are verified and normalised into traceable dataset fields. Across four initial papers, the pipeline achieved an aggregate precision of 99%. Further evaluation will assess F1 and compare the pipeline with a single-agent pipeline.

Machine Learning-based analysis of data sets from Graphene Transistor Biosensors

Track: Emerging Research

Authors: Muhammad Hamza Rafique Bhatti (University of Plymouth, Plymouth, United Kingdom), Asiya Khan (University of Plymouth, Plymouth, United Kingdom), Shakil Awan (University of Plymouth, Plymouth, United Kingdom), Dena Bazazian (University of Plymouth, Plymouth, United Kingdom)

Presenter: Muhammad Hamza Rafique Bhatti (University of Plymouth, Plymouth, United Kingdom)

Abstract: This paper proposes a systematic Machine Learning (ML) pipeline to analyze Graphene Field-Effect Transistor (GFET) biosensor data. The primary objective of this study is to validate the effectiveness of each biosensor stage during the fabrication and characterization of the GFET biosensor. Each biosensor stage was investigated by predicting the rate of change of resistance using GFET electrical measurement data. The GFET biosensor data were preprocessed by handling missing values and removing outliers. This study utilized the Gradient Boosting (GB) algorithm to predict changes in resistance during GFET development. The GB algorithm achieved the highest predictive performance, with an $R2$ of 0.9984, RMSE of 51.19, and MAE of 36.02. The proposed approach effectively integrates ML techniques with biosensor data to enable stage-wise validation of GFET development.

Domain-Aware Deep Learning for Forest Carbon Mapping: Interpretability, Causal Constraints, and Learning Under Biased Supervision

Track: Emerging Research

Authors: Reza M. Asiyabi (School of Geosciences, The University of Edinburgh and National Centre for Earth Observation (NCEO)), The SEOSAW Partnership (School of Geosciences, The University of Edinburgh), Steven Hancock (School of Geosciences, The University of Edinburgh and National Centre for Earth Observation (NCEO)), Casey M. Ryan (School of Geosciences, The University of Edinburgh)

Presenter: Reza M. Asiyabi (School of Geosciences, The University of Edinburgh and National Centre for Earth Observation (NCEO))

Abstract: Mapping forest aboveground biomass (AGB) from satellite imagery is essential for global carbon accounting, yet the data conditions that underpin it are among the most hostile in applied machine learning: labels are sparse, drawn from structurally incompatible instruments, missing not at random, and governed by biophysical processes that black-box models neither encode nor respect. We present two complementary frameworks, PG-CBM and StruMPL, that address these challenges jointly, and argue that embedding domain knowledge into AI architecture is a prerequisite for trustworthy AI in high-stakes environmental applications.

The IDEA of Us: An Identity-Aware Architecture for Autonomous Systems

Track: World Class Research Track

Authors: Amel Bennaceur (The Open University), Carlos Gavidia-Calderon (The Alan Turing institute), Anastasia Kordoni (Lancaster University), Mark Levine (Lancaster University), Bashar Nuseibeh (The Open University)

Presenter: Amel Bennaceur (The Open University)

Abstract: Autonomous systems, such as drones and rescue robots, are increasingly used during emergencies. They deliver services and provide situational awareness that facilitate emergency management and response. To do so, they need to interact and cooperate with humans in their environment. Human behaviour is uncertain and complex, so it can be difficult to reason about it formally. In this article, we propose IDEA: an adaptive software architecture that enables cooperation between humans and autonomous systems, by leveraging the social identity approach. This approach establishes that group membership drives human behaviour. Identity and group membership are crucial during emergencies, as they influence cooperation among survivors. IDEA systems infer the social identity of surrounding humans, thereby establishing their group membership. By reasoning about groups, we limit the number of cooperation strategies the system needs to explore. IDEA systems select a strategy from the equilibrium analysis of game-theoretic models that represent interactions between group members and the IDEA system. We demonstrate our approach using a search-and-rescue scenario, in which an IDEA rescue robot optimises evacuation by collaborating with survivors. Using an empirically validated agent-based model, we show that the deployment of the IDEA system can reduce median evacuation time by 13.6%.

Human–AI Complementarity Through Difference: Expanding Decision Attributes and Integrating Divergent Evaluations in Everyday Decisions

Track: Emerging Research

Authors: Adrian Banks (University of Surrey), David Gamblin (Birkbeck College, University of London)

Presenter: Adrian Banks (University of Surrey)

Abstract: AI decision support can improve human judgement, but human–AI systems rarely outperform both human and AI alone. We tested whether large language models can produce such synergy in everyday multi-attribute decisions. Participants reported a current low-stakes decision involving two options. We compared four evaluations: an AI-only rating, an AI-only multi-attribute utility model, a human-only multi-attribute utility model, and a fully augmented model combining human- and AI-generated attributes with AI-suggested ratings and final human selection. One week later, participants rated the success of the decision outcome. The fully augmented model predicted later outcome success more accurately than the human-only, AI-only, and AI-only multi-attribute models. Augmentation also increased the number of attributes considered, and greater disagreement between human and AI evaluations predicted larger gains from augmentation. These findings suggest that human–AI synergy can emerge when AI broadens the information considered by introducing different perspectives while humans retain control over personally relevant evaluation.

Self-supervised Learning of Imperfect Crystal Embeddings

Track: Emerging Research

Authors: Sohan Seth (University of Edinburgh), Dagmara Panas (University of Edinburgh), James Cumby (University of Edinburgh)

Presenter: Sohan Seth (University of Edinburgh)

Abstract: Technologically useful crystalline materials frequently possess complex compositions which give rise to imperfect (substitutionally disordered) structures. Despite this, most existing ML models cannot handle disorder. Here we present Wyckoff-orbit multigraph (\textsc{WORM}); a symmetry-aware, natively disordered graph architecture trained using self-supervised learning. Initial results demonstrate that the proposed architecture encodes chemistry relevant knowledge, and readily differentiates between disordered compounds adopting very similar crystal structures.

Toward Human Rights Benchmarking for LLMs: A Pilot Methodology

Track: World Class Research Track

Authors: Wm. Matthew Kennedy (Oxford Internet Institute, University of Oxford / King's College London), Savannah Thais (Hunter College)

Presenter: Wm. Matthew Kennedy (Oxford Internet Institute, University of Oxford / King's College London)

Abstract: Large language models (LLMs) increasingly mediate decisions that affect what human rights are realized, and how. Yet, no evaluation benchmark exists to assess whether they can reason correctly about human rights. To this end, we report our efforts to develop a robust and scalable methodology for creating HumRightsBench---the first expert-validated, scenario-based benchmark for human rights reasoning. We adapt the IRAC framework for legal reasoning to better suit the unique reasoning patterns of human rights work (substituting P, "proposing remedies," for C, "legal conclusion," yielding IRAP) to structure our evaluation heuristics. We also produce a pilot series of authentic scenarios designed to implicate the many dimensions of real-world human rights issues and annotated by human rights lawyers and professionals across the world. Ultimately, we find that frontier model accuracy scores range considerably across legal reasoning tasks (0.339-0.577), which strongly implies that HumRightsBench is a capable instrument for advancing this emerging subfield of AI evaluations science at a critical time.

Can LLMs work in Reflexive Thematic Analysis?

Track: Emerging Research

Authors: Bethany Ahluwalia (University of Southampton), Adriana Wilde (University of Southampton), Daniel Ashton (University of Southampton), Richard Gomer (University of Southampton)

Presenter: Bethany Ahluwalia (University of Southampton)

Abstract: The integration of Large Language Models (LLMs) into qualitative research has sparked significant debate, particularly in terms of their capacity for reflexivity, generally understood to be a researcher's active engagement in self-interrogation of their own subjectivity and context. Rather than entirely rejecting the use of LLMs in thematic analysis, we find potential to reframe their role through the lenses of distributed cognition and dialogic spaces. The use of LLMs as automated Socrates or Devil's advocates can enable researchers to harness the friction generated by the model's otherness to challenge their own assumptions and enhance their reflexive practices. This is a shorter version of our position paper at the Synergy workshop at AVI 2026, which develops these arguments in greater depth.

Deepfakes and Social Legibility

Track: Emerging Research

Authors: Guangwei Teng (University of York, Department of Philosophy), Tom Stoneham (University of York, SAINTS UKRI CDT)

Presenter: Tom Stoneham (University of York, SAINTS UKRI CDT)

Abstract: Before generative imaging, realistic photographs provided social legibility, a readily available and reliable resource for society, something that social norms permitted an appeal to as evidence without checking credibility. Authentication is widely expected to revive the default trust in realistic images that deepfakes have undermined: as provenance and detection mature, the thought goes, a realistic image will again be believable on sight. Laas's blockchain-backed source-authentication [10] is one such proposal. We argue that no authentication protocol can restore that lost warrant, for two structural reasons. First, the lost default was a type-level entitlement not to check realistic images, and verifying is the act it licensed one to skip: per-item authentication restores checking, never the entitlement. Second, an image's credential is separable from the image content (a screenshot detaches it), as a TLS certificate welded to its connection is not, so images themselves are not believable on sight. That the default has lapsed is not news; the structural ceiling on repairing it is. The ceiling splits the engineering problem: capture-integrity is cheap and automatable, but evidential standing—a party answerable for what an image represents—is scarce, unautomatable, and useless unless legible to the ordinary public.

A System-Level Systolic Array Accelerator for Convolution Workloads

Track: Emerging Research

Authors: Eleni Tselepi (University of Edinburgh), Cristian Sestito (University of Edinburgh), Shady Agwa (University of Edinburgh), Themis Prodromakis (University of Edinburgh)

Abstract: Vision generative AI models rely heavily on convolutions and matrix multiplications. As these models grow in complexity in pursuit of higher accuracy, their computational and memory demands increase significantly, making deployment on edge devices increasingly difficult. This work presents a complete system-level specialised hardware architecture based on systolic arrays. By co-designing the compute array and the memory subsystem together, the proposed architecture makes vision generative AI models accessible on edge devices, unlocking their use across a wider range of platforms and real-world applications.

Towards AI-Assisted Air Traffic Control: Learning Navigation and Obstacle Avoidance Rules with Transformers

Track: Emerging Research

Authors: James Edwards (University of Exeter), George De Ath (University of Exeter), Richard Everson (University of Exeter)

Presenter: James Edwards (University of Exeter)

Abstract: The sequential nature of air traffic control, in which a controller issues successive instructions to navigate aircraft safely to their destinations, closely mirrors the structure of the sequence modelling problems at which transformer architectures excel. We investigate whether a decoder-only transformer, built from the same components as modern large language models, can learn goal-directed navigation and obstacle avoidance rules directly from demonstration data. Working in a simplified two-dimensional sector, we train a model on synthetic expert trajectories in which a controlled aircraft transits the sector while avoiding stationary traffic, issuing one of a discrete set of compass headings at each step. Run closed-loop on traffic configurations unseen during training, the model generates coherent, efficient trajectories that largely respect the required separation from other aircraft, suggesting that the architectural advances driving progress in language modelling could transfer to the sequential decision problems of air traffic control.

Permutation Equivariant Neural Networks for Antisymmetric Tensors

Track: Emerging Research

Authors: Edward Pearce-Crump (Imperial College London)

Presenter: Edward Pearce-Crump (Imperial College London)

Abstract: Antisymmetric tensors change sign under index swaps, yet learning methods that exploit this structure remain largely unexplored. We characterise all linear permutation equivariant maps between antisymmetric power spaces of Rn, showing that they are determined by a class of objects that we term antispherical bipartitions. We further introduce IO patterns, a memory-efficient representation that eliminates the need to construct and store large weight matrices. This framework is well-suited to graph-structured data with signed or directed edges, where antisymmetry encodes the orientation of edges and permutation equivariance encodes the symmetry of node relabellings. We evaluate our approach on the Wiki-RFA dataset, showing improved performance over a baseline MLP, and demonstrate strong scalability and robustness to noise on synthetic graphs of varying size and noise levels.

Geo-SMILE: Enhancing Explainability in Geospatial Machine Learning through Spatial Awareness

Track: Emerging Research

Authors: SHADIEH MOHAMMADI KAZAJ (university of hull)

Abstract: This paper introduces Geo-SMILE, a spatially aware explainability framework for geospatial machine learning. Building on the SMILE methodology, Geo-SMILE combines spatially structured perturbations with local surrogate modelling to generate interpretable feature- and location-based explanations. The framework addresses key GeoAI challenges, including spatial autocorrelation, regional heterogeneity, and black-box model opacity, by treating geographic location as an integral part of the explanation process. Using evaluation criteria such as fidelity, faithfulness, stability, and consistency, Geo-SMILE assesses explanation quality in spatial settings. It offers a transparent and flexible approach for applying explainable AI in urban analytics and other location-sensitive domains.

Automating FPGA Accelerator Design using Large Language Models

Track: Emerging Research

Authors: Vinamra Sharma (University of Glasgow), Xingjian Fu (University of Glasgow), Jude Haris (University of Glasgow), José Cano (University of Glasgow)

Abstract: The growing complexity of FPGA accelerator design presents a major challenge for researchers and developers, requiring extensive exploration of architectural trade-offs involving computation, memory organization, and resource utilization in order to design an efficient accelerator design. Although recent advances in Large Language Models (LLMs) have demonstrated strong capabilities in reasoning and code generation, their potential for guiding hardware design workflows remains underexplored. This short paper presents an emerging research direction towards self-improving FPGA accelerator design through the integration of LLM-guided reasoning and iterative hardware feedback. Building upon the SECDA ecosystem, we explore how retrieval-augmented reasoning, structured design space exploration, and hardware evaluation data can be combined to support automated accelerator design generation and refinement. Preliminary results demonstrate the successful generation and execution of three accelerator designs, providing early evidence that hardware feedback can be leveraged to improve future design decisions. This work highlights the potential of AI-assisted hardware co-design and outlines a path towards more autonomous FPGA accelerator development workflows.

AI for Interpretable Model Discovery in the Natural Sciences

Track: Emerging Research

Authors: Dabin Kwon (UCL), Reilly Tilbury (UCL), Rajah Nutakki (UCL), Kevin Miller (Google DeepMind, UCL), Kim Stachenfeld (Google DeepMind, Columbia University), Kenneth Harris (UCL)

Abstract: We present EDGAR (Equation Discovery with Graphical AI Reasoning), a general-purpose equation discovery engine. Building on LLM-driven evolutionary algorithms such as AlphaEvolve, EDGAR uses automated model evaluation and cross-validation to discover mathematical models for arbitrarily complex datasets. Applying this system to neural recordings from the mouse visual cortex, we discovered a novel single-cell tuning function that both improves the fit to experimental data and explains the high dimensionality of population codes. Unlike AlphaEvolve, EDGAR is open-source, and includes innovations such as a parameter-estimation process that makes evaluation more efficient and reduces compute cost (\$8.25 per iteration on the neural dataset) - an important aspect which disproportionately affects the adoption of AI-assisted research in academic labs.

A Framework for Designing Sociotechnical Evaluation Specifications for LLMs in High-Stakes Domains

Track: Emerging Research

Authors: Maria Waheed (University of Nottingham), Aislinn Bergin (University of Nottingham), Emily Thorn (University of Nottingham), Teresa Castle-Green (University of Nottingham), Sebastian Löbbers (Queen Mary University of London), Mahmud Elahi Akhter (Queen Mary University of London), Joshua Kelsall (University of Warwick), Tom Sorell (University of Warwick), Robert Procter (University of Warwick), Maria Liakata (Queen Mary University of London)

Abstract: Large Language Models (LLMs) are increasingly deployed in high-stakes domains such as healthcare and law, yet current evaluation practices struggle to account for what happens in the real-world. Assurance is essential for responsible AI development and deployment, and evaluation is a key component of any assurance process. We argue that responsible LLM evaluation must be grounded in real-world needs and sociotechnical in nature. Requirements need to be gathered from stakeholders within these domains so that evaluations appropriately reflect what is needed within specific contexts. We present a three-component framework structured around requirements, aspects, and metrics, along with a methodology to support operationalisation of responsible sociotechnical evaluation of LLMs. Together, the framework and methodology support the development of task and domain specific evaluation packages that are traceable back to co-produced stakeholder requirements, offering a structured approach to generating meaningful evidence for AI assurance processes.

A Dual-Edge Spatial–Jacobian Image Graph for Interpretable Diabetic Retinopathy Grading

Id: 13301 - UKAIRS26-197

Submitted at: 22/06/2026 20:16:21

Track: Emerging Research

Paper status: Accepted

Authors: Inam Ullah (University of Southampton)

Presenter: Inam Ullah (University of Southampton)

Abstract: Automated diabetic retinopathy (DR) grading from colour fundus photographs can achieve strong predictive performance, but clinical interpretation requires more than an image-level label. It requires understanding how lesion evidence is distributed around retinal vessels and how this evidence relates to quantitative vascular biomarkers. We present a dual-edge spatial–Jacobian image graph for interpretable DR grading. Each fundus image is represented as a graph node with four aligned evidence streams: AutoMorph vessel information ($X1$), DR-XAI-style lesion evidence maps ($X2$), a 128-dimensional lesion-based contrastive image embedding ($X3$), and AutoMorph morphometric biomarkers ($X4$). The spatial edge branch ($X12$) encodes vessel–lesion geometry, while the Jacobian branch ($X34$) models embedding–biomarker sensitivity. Lightweight two-token attention fuses both edge families into a final image graph. On 2,910 matched non-augmented APTOS images, the full graph achieves 0.8076 accuracy, 0.8312 quadratic weighted kappa, 0.5915 macro-F1, and 0.9330 adjacent-grade accuracy; referable DR reaches 0.9055 accuracy and 0.9711 AUROC. The framework is positioned as an explainable representation- learning tool for lesion–biomarker hypothesis generation, rather than as a deployment-ready clinical classifier. The code is available at https://github.com/Inamullah-Colab/dual-edge-dr-graph-xai.

Welfare-Aware EV Charging Dispatch with Vehicle-to-Grid and Storage

Track: Emerging Research

Authors: Fariba Dehghan (University of Southampton), Sebastian Stein (University of Southampton), Vahid Yazdanpanah (University of Southampton), Stephanie Gauthier (University of Southampton)

Abstract: This paper develops a day-ahead mixed-integer linear programming framework for coordinating an islanded photovoltaic-powered electric vehicle charging station with stationary storage, backup generation, and vehicle-to-grid capability. The model co-optimises EV admission, charger assignment, charging and discharging, storage operation, photovoltaic utilisation and curtailment, and backup generation to maximise social welfare under infrastructure and energy-resource constraints. The framework is evaluated over a 24-hour horizon using bootstrapped empirical charging-session data and three station archetypes: a PV-rich workplace hub, a PV-balanced public hub, and a PV-scarce rapid-DC hub. Preliminary results show that welfare and completion objectives align when resources are abundant, but diverge in constrained public and rapid-DC settings, where welfare-oriented scheduling increases total welfare while serving fewer requests. These findings show that maximising throughput is not always equivalent to maximising system value, and provide a physically feasible benchmark for future scalable and online EV charging management under renewable uncertainty.

Continual Generalized Category Discovery via Enhancing Representation Learning

Track: Emerging Research

Authors: Jianqiao Weng (Cardiff University), Okatay Karakus (Cardiff University), Yuhua Li (Cardiff University)

Abstract: This work investigates continual generalized category discovery, where learning systems are required to preserve reliable recognition of previously observed categories while progressively discovering and incorporating novel categories from evolving unlabeled data streams. In this work, a unified exemplar guided framework is proposed for jointly leveraging central exemplars to preserve representative category characteristics and peripheral exemplars to capture discriminative boundary information. The proposed method regularizes the representation space through similarity constraints defined with respect to category prototypes and introduces an adaptive novelty estimation strategy for identifying observations from potentially unseen categories. Preliminary results on CIFAR 100 demonstrate that the proposed approach achieves effective recognition of known categories and exhibits promising capability in discovering and integrating novel categories. Future works intend to broaden the empirical evaluation on large-scale benchmarks, extensive and comprehensive comparisons, and systematic ablation analysis.

L2-Bench: An Evaluation Benchmark for Measuring LLM Capabilities in Second Language Education

Track: Emerging Research

Authors: James Edgell (Oxford University Press), Matthew Kennedy (Oxford Internet Institute)

Presenter: James Edgell (Oxford University Press)

Abstract: Despite rapid AI adoption in education, rigorous evaluation of AI-powered educational (AIED) systems remains critically underdeveloped, particularly in second language (L2) education, one of the most common yet least evaluated AI applications. We introduce L2-Bench, an open-source benchmark of 1,000+ task-response pairs to aid the pedagogy-led evaluation of LLM capabilities relating to language learning and assessment. Crucially, L2-Bench measures model performativity on the application of learning experience design principles rather than mere knowledge of those principles or broad learning outcomes. Our contributions include: (1) a validated taxonomy of 12 competencies and 31 subcompetencies validated by 200+ expert practitioners (task authenticity = 4.42/5.00, criteria adequacy = 4.18/5.00); (2) a rubric-based evaluation methodology that we believe can, if adapted, generalize to similar (open-ended, qualitative) disciplines; (3) an evaluation dataset that produces reliable signal about model strengths, weaknesses, and contextual robustness across diverse L2 education scenarios. We find that, among large models, Claude Opus 4.7 performs best overall (85.5%), though is marginally outperformed on several constituent tasks. We also find that performance drops notably on harder tasks $\in$ (69.9%–73.4%). L2-Bench provides education stakeholders better methods to make more informed decisions about real-world AIED adoption, use, and governance, while advancing the maturing science of AI evaluations for education.

A Principled Probabilistic Machine Learning Framework for Optical Communication Channels

Track: Emerging Research

Authors: SINA SEYFADINI KUHBANANI (Aston University), David Saad (Aston University), Sergei K Turitsyn (Aston University), FELIPPE ALVES (TELUS Digital Research Hub)

Presenter: SINA SEYFADINI KUHBANANI (Aston University)

Abstract: Optical communication networks carry over 90% of global internet traffic and are therefore central to modern society and economy. Quadrature Amplitude Modulation (QAM) is a way of sending digital data by changing both the amplitude and the phase of a carrier signal. QAM maps groups of bits onto symbols that are points in a two dimensional grid, where on the complex plane the horizontal axis is called the in-phase component I and the vertical axis is called the quadrature component Q, a transmitted symbol is therefore a complex number of the form s=I+jQ. The observed constellation sample after transmission generally appears as a perturbed version of the originally transmitted symbol, since channel impairments distort its position in the complex plane due to the sources of noise and inter-symbol-interference (ISI), a form of signal distortion in digital communications where a transmitted symbol spreads out in time and overlaps with subsequent symbols. In this work, we develop a principled probabilistic machine learning inference framework for dispersive channels. Our approach combines Generalised Belief Propagation (GBP) and a Viterbi algorithm jointly with a Gaussian-mixture model of inter-symbol interference. Large-scale transmission data are generated using VPIphotonics, a well-established commercial optical system simulation environment, across different symbol sequences and launch powers. Using 2^{25} transmitted symbols, the model is trained on a statistically rich dataset that supports reliable estimation of sequence dependent error patterns. The model is then incorporated into optical transmission procedures, leading to a significant Mutual-Information (MI) gain and Bit Error Rate (BER) reduction.

Investigating the Specific Cultural Proxies That People Need From Text-To-Image Models

Track: Emerging Research

Authors: Boyd Migisha (Swansea University)

Abstract: Generative Artificial Intelligence (GenAI) is commonly used to create realistic text or images from human input. These models have been shown to have several deficiencies (biases) present in their outputs. One such bias is the scarcity or absence of cultural information in their outputs when inputs are of non-Western cultures. We want to help close this gap by investigating Text To Image (TTI) models to understand (a) what specific cultural proxies people rely on to judge the cultural appropriateness of TTI images, and (b) how well or poorly these proxies are rendered by the TTI models. The results of this study would (a) begin to create a taxonomy of cultural proxies relevant to the generation of purposeful images in non-Western settings, (b) highlight which proxies TTI models struggle to reproduce, and (c) discuss how to close the gap between the people's needs and models' capabilities.

AI based Mediation for Collaborative Decision Making under Hidden Profile Conditions

Track: Emerging Research

Authors: Thiara de Alwis (University of Bristol), Kira Clements (University of Bristol), Paul Marshall (University of Bristol)

Abstract: Decision making in group contexts often suffers from process failures such as bias and the effects of interpersonal dynamics. As AI use has increased in recent years, so too has its integration into social situations and conversational settings. This project brings together psychological research and advances in AI to implement an LLM based mediator to facilitate small group discussions, particularly in the context of asymmetric information conditions - a Hidden Profile. The Hidden Profile paradigm provides an experimental condition to explore dynamics such as shared information and initial preference bias by creating a decision making scenario in which the optimal choice can only be reached through effective information sharing and evaluation. This project utilises over 40 years of hidden profile literature to create an intervention framework operationalised through a multi agent system making use of observer agents to track conversational state. Findings indicate the utility of such a mediation system for lower performing groups and deployment at scale.

Beyond Thresholds: Learning Perioperative Risk from Intraoperative Physiological Trajectories

Track: World Class Research Track

Authors: Sadaf Sohrabi (Imperial College London), Shehan Hettiaratchy (Imperial College Healthcare NHS Trust), Aldo Faisal (Imperial College London)

Presenter: Sadaf Sohrabi (Imperial College London)

Abstract: Clinical risk scores for surgical patients are computed once from preoperative variables and consequently fail to capture the physiological dynamics that unfold during surgery. We ask whether intraoperative vital-sign time series, routinely recorded but rarely used for real-time risk stratification, provide predictive information that static models fail to capture. Using 130,960 operations from the INSPIRE perioperative dataset, we compare four modelling arms across progressively richer feature regimes (static baseline, physiological summaries, HMM-derived trajectory features, and their combination) and evaluate performance across four postoperative outcomes using grouped cross-validation. We find that learned trajectory representations capture complementary prognostic structure to threshold-based summaries for ICU-related outcomes. Predictive signal is detectable as early as 15 minutes after anaesthesia onset, and gains are largest in patients whose static risk profile underestimates true intraoperative vulnerability. These results demonstrate that temporal modelling of multivariate intraoperative time series provides clinically meaningful information that threshold-based approaches systematically miss. This paper was accepted as a poster at the Structured Data for Health (SD4H) Workshop at ICML 2026.

Compressing Quaternion Convolutional Neural Networks for Audio Classification

Track: World Class Research Track

Authors: Arshdeep Singh (King's College London and The Gen AI Hub (The AI Hub in Generative Models)), Vinayak Abrol (IIIT Delhi, India), Mark D. Plumbley (King's College London and the Gen AI Hub (The AI Hub in Generative Models))

Abstract: Conventional Convolutional Neural Networks (CNNs) in the real domain have been widely used for audio classification. However, CNNs have limited ability to capture correlations across channels when processing multi-channel inputs. This can lead to suboptimal feature learning, particularly for complex audio patterns such as multi-channel spectrogram representations. Quaternion Convolutional Neural Networks (QCNNs) address this limitation by employing quaternion algebra to jointly capture inter-channel dependencies, enabling more compact models with fewer learnable parameters while better exploiting the multi-dimensional nature of audio signals .However, QCNNs exhibit higher computational complexity due to the overhead of quaternion operations, resulting in increased inference latency and reduced efficiency compared to conventional CNNs, posing challenges for deployment on resource-constrained platforms. To address this challenge, this study explores knowledge distillation (KD) and pruning, to reduce the computational complexity of QCNNs while maintaining performance. Our experiments on audio classification reveal that pruning QCNNs achieves similar or superior performance compared to KD while requiring less computational effort. Compared to conventional CNNs and Transformer-based architectures, pruned QCNNs achieve competitive performance with a reduced learnable parameter count and computational complexity. On the AudioSet dataset, the pruned QCNN14 achieves performance comparable to that of the conventional CNN14, while using only 10% of the parameters and 70% of the computational complexity of conventional CNN14. Furthermore, pruned QCNNs generalize well across multiple audio classification benchmarks, including GTZAN for music genre recognition, ESC-50 for environmental sound classification and RAVDESS for speech emotion recognition.

Incremental Semantic Processing for Dialogue: Neuro-Symbolic Grammar Induction and Mechanistic Evaluation

Track: Emerging Research

Authors: Arash Ashrafzadeh (Heriot-Watt University)

Abstract: Dialogue understanding requires constructing meaning of utterances word by word in real time, similar to how humans do it, yet current models for conversational AI often lack this incremental semantic processing. This proposal motivates the importance of incrementality, and reviews some of the work on incremental NLU and suggests potential paths for future research, focussing on cognitively-plausible wide-coverage semantic parsing through neuro-symbolic grammar induction in Dynamic Syntax with Type Theory with Records (DS-TTR), alongside mechanistic interpretability methods for studying incrementality in language models.

Identifying and Classifying Peculiar Galaxies in the COSMOS-Web Field

Track: Emerging Research

Authors: Kai Madgwick (University of Manchester)

Presenter: Kai Madgwick (University of Manchester)

Abstract: A wide range of galaxies can be classed as being "Peculiar", or possessing unusual visual morphologies. These include irregular, tidally disturbed, and merging galaxies. Individually, these groups are difficult to identify with supervised Machine Learning models due to their rarity combined with a lack of accurate labels. Using a CNN classifier trained on JWST morphologies of galaxies in the CEERS field, we identify a general population of 6020 Peculiar galaxies in a subset of the larger COSMOS-Web field. A VAE was trained on this set of galaxies to extract image features, following which unsupervised Hierarchical Clustering was employed to identify groups of galaxies. Visual inspection of the UMAP showed several of these groups to be outliers, and a combination of image and morphological parameter analysis revealed these galaxies to be predominantly merging systems, hence demonstrating that mergers can be identified without precise prior labels in the training data.

Comparing TimeGAN and Diffusion-TS for Rare-Event Indoor Air Quality Data Augmentation

Track: Emerging Research

Authors: Lewis Hall Rudford (Swansea University), Sara Sharifzadeh (Swansea University), Nicholas Micallef (Swansea University), Paul Lewis (Raven Delta Group), Daniele Doneddu (Swansea University)

Presenter: Lewis Hall Rudford (Swansea University)

Abstract: Generative adversarial networks and diffusion models provide contrasting approaches to synthetic time-series generation; however, their relative value for imbalanced indoor air quality forecasting tasks remains unclear. This study presents a controlled comparison of TimeGAN- and Diffusion-TS-based augmentation for rare-event indoor air quality forecasting. We compare TimeGAN, Conditional TimeGAN, Diffusion-TS, and Conditional Diffusion-TS within a common forecasting and evaluation framework. Each method generates multivariate 36-step sequences comprising 24 historical observations and 12 forecast targets, with conditional variants explicitly targeting windows containing elevated CO2 concentrations. Synthetic samples are added only to the training data, while performance is evaluated on a chronologically separated real test set using overall and elevated-regime root mean squared error. The comparison reveals trade-offs between the techniques as TimeGAN achieves the lowest overall forecasting error, Diffusion-TS provides a stronger balance between overall and elevated-regime performance, and Conditional Diffusion-TS produces the largest reduction in error during elevated CO2 conditions. These results show that generator choice affects the balance between average-case accuracy and rare-event sensitivity, with diffusion-based methods becoming increasingly advantageous as performance in the minority regime is prioritised.

Trust Thy Neighbours: Probabilistic Pseudo-Labelling via Knowledge Distillation for Semi-Supervised Wildlife Monitoring

Track: Emerging Research

Authors: Rachael Laidlaw (University of Bristol), Dan Morris (Google Research), Sara Beery (Massachusetts Institute of Technology), Michael Wray (University of Bristol)

Presenter: Rachael Laidlaw (University of Bristol)

Abstract: Training computer vision models for biodiversity monitoring tasks demands a vast number of labelled images from the target ecosystem. While unlabelled camera trap data is globally abundant, the expert effort needed for accurate annotations creates a critical bottleneck, and pre-existing classifiers often struggle due to label space mismatch and poor generalisability. We therefore propose a model-agnostic and ecosystem-agnostic knowledge-distillation approach to tackle this labelling dilemma. Rather than relying on self-ensembling baselines like Mean Teacher, our method leverages the robust, general-purpose representations from a foundation model (DINOv2). We probe its latent space using unlabelled target data, mapping the resultant embeddings to the required label environment via a sparse support set of known anchor points. These pseudo-labelled images are then incorporated into local model training without any extra manual labour. Evaluated on the Idaho Camera Traps dataset, our method yields a boost of over 10% in overall accuracy for few-label regimes relative to baselines omitting unlabelled data. Crucially, using just 20% real labels, our framework achieves performance comparable to the 100% fully-supervised upper bound, vastly reducing the human annotation necessary.

Forest Gain Foundations: Earth observation embeddings for large-scale forest gain typology mapping

Track: Emerging Research

Authors: Ce Zhang (University of Bristol), Nantheera Anantrasirichai (University of Bristol), James Brock (University of Bristol)

Presenter: James Brock (University of Bristol)

Abstract: This paper presents an ongoing study into the use of Earth observation foundation model embeddings for large-scale forest gain detection and typology mapping across Europe. We introduce a novel dataset pairing pixel-level forest gain masks with ForTy-derived typology confidence distributions, and evaluate Alpha Earth and TESSERA embeddings against raw Sentinel-1/2 imagery under a common weakly supervised multi-task framework. The work addresses a critical gap in open, operationally useful forest gain monitoring tools, with direct relevance to policy commitments such as the EU Forest Strategy 2030.

Generative AI Inference for Consensus in Blockchains

Track: Emerging Research

Authors: Wan Ki Wong (University of Edinburgh), Michele Ciampi (University of Edinburgh), Rik Sarkar (University of Edinburgh)

Abstract: Blockchains are energy hungry and unsustainable because they employ repeated hash computations as a way to slow down and avoid simultaneous block creation by multiple parties. Instead, we propose use of Generative AI computations as a way to slow down block creation. In our system called Proof-of-ML, query makers submit generative AI queries, and service providers (block miners) compute their responses, and get the right to create a block after sufficient work. Integrity of the system is maintained via cryptographic systems such as zero knowledge proofs of generative models. We have developed protocol, proofs and implementations of this system. Experiments show that our system provides steady performance. This system simplifies participation of small service providers in AI ecosystems. It improves sustainability by using a common sequence of computations for Generative AI and Blockchain regulation.

An Informed Design to Explore Psychological safety of Artificial Intelligence for Older Adults: A Systematic Review, Expert Consultations and Public Engagement.

Track: Emerging Research

Authors: Prenika Anand (University of York), Peter Knapp Knapp (University of York), Lina Gega (University of York), Nick Pears (University of York)

Presenter: Prenika Anand (University of York)

Abstract: Assisted Living Technologies (ALTs) including those led by Artificial Intelligence (AI) are being explored for their potential in supporting health, mobility and social connectedness of older adults. While aimed at fostering safety, independence and wellbeing, the evidence on the safety of such technologies is inadequate and fraught with ethical challenges. The concepts of psychological harm and safety in this field of human AI interaction have specifically been underexplored. While traditional safety methodologies focus primarily on physical safety, this research study explores the potential psychological harms posed by human-AI interactions involving older adults. Drawing on a systematic review of existing evidence in the field, expert consultations through a research podcast, and a primary survey, this research aims to propose a theoretical framework to conceptualize psychological harm and safety. Further, it aims to undertake an empirical application of this framework to a real-world scenario involving older adults and AI. The study leverages public engagement to inform its design and conduct, and gain perspectives on its relevance to older adults. This approach also aligns with research guidance on incorporating the lived experiences of end users in the design of scientific evaluations of AI. The beneficiaries of this project are older adults, caregivers, health professionals, policy makers, product developers and academics in social care technologies. Overall, the research aims to build on the existing limited research on psychological harms of AI, promote a wider discourse on safer deployment of AI for older people, and contribute to future research, design and regulatory practices in this area.

Strategic Verification in Agentic Escrow: A Game-Theoretic Analysis of ERC-8183

Track: Emerging Research

Authors: Yanzhen Li (University of Bristol), Zining Wang (University of Bristol)

Abstract: Agentic escrow lets autonomous clients fund jobs, providers submit work, and evaluators trigger settlement. This architecture, exemplified by ERC-8183, makes execution programmable but shifts trust to post-agreement verification. We model client-provider-evaluator escrow as a dynamic game with optional hooks and derive compact incentive conditions for provider effort, evaluator verification, and signal-truthful reporting. The analysis shows a negative result: without audit, stake, reputation, or equivalent liability, evaluator-based escrow is not robust to settlement-changing side payments. Hooks can improve incentives, but only through the concrete channels they affect: enforceable collateral, auditability, reputation linkage, and liveness.

Soft law approaches to Agentic AI: Challenges and Future directions

Track: Emerging Research

Authors: Myrto Tachia (SAINTS, University of York)

Abstract: [Poster presentation] The paper will be exploring some recent developments of non-binding regulatory instruments (soft law) specifically related to Agentic AI.

Graph Bridge Diffusion for Scaffold Decoration

Track: Emerging Research

Authors: Magdalena Proszewska (University of Edinburgh)

Abstract: Fragment-constrained molecular generation---completing a partial molecule around a fixed scaffold---is a central task in structure-based drug discovery, yet existing graph diffusion models handle it only at inference time, leaving the model blind to the constraint during training. We present a graph diffusion framework that treats scaffold conditioning as an integral part of training rather than a post-hoc fix. Concretely, we cast scaffold decoration as a graph-bridge problem: a variational scaffold encoder produces a stochastic latent that is injected into every denoising step, and scaffold nodes are held fixed throughout both training and sampling. This keeps the model in-distribution when constraints are applied at inference, and allows a single unified model to support both conditional and unconditional generation via dropout. On the MolGenBench scaffold-decoration benchmark, our method achieves the best uniqueness and diversity among graph-based models, while relying on orders of magnitude less training data than competing sequence-based approaches.

Stuart-Landau Oscillatory Graph Neural Network

Track: World Class Research Track

Authors: Kaicheng Zhang (University of Edinburgh), David Reynolds (University of Warsaw), Piero Deidda (Scuola Normale Superiore), Francesco Tudisco (University of Edinburgh)

Presenter: Francesco Tudisco (University of Edinburgh)

Abstract: Oscillatory Neural Networks (ONNs) are an emerging class of physics-inspired architectures designed to mitigate oversmoothing and vanishing gradient problems in deep NNs. In this work, we introduce the Complex-Valued Stuart-Landau Graph Neural Network (SLGNN), a novel architecture grounded in Stuart-Landau oscillator dynamics. Stuart-Landau oscillators are canonical models of limit-cycle behavior near Hopf bifurcations, which are fundamental to synchronization theory and are widely used in e.g. neuroscience for mesoscopic brain modeling. Unlike harmonic oscillators and phase-only Kuramoto models, Stuart-Landau oscillators retain both amplitude and phase dynamics, enabling rich phenomena such as amplitude regulation and multistable synchronization. The proposed SLGNN generalizes existing phase-centric Kuramoto-based OGNNs by allowing node feature amplitudes to evolve dynamically according to Stuart-Landau dynamics, with explicit tunable hyperparameters (such as the Hopf parameter and the coupling strength) providing additional control over the interplay between feature amplitudes and network structure. We conduct extensive experiments across node classification, graph classification, and graph regression tasks, demonstrating that SLGNN outperforms existing OGNNs and establishes a novel, expressive, and theoretically grounded framework for deep oscillatory architectures on graphs.

Neural-HSS: Hierarchical Semi-Separable Neural PDE Solver

Track: World Class Research Track

Authors: Pietro Sittoni (Gran Sasso Science Institute), Emanuele Zangrando (Gran Sasso Science Institute), Nicola Guglielmi (Gran Sasso Science Institute), Francesco Tudisco (University of Edinburgh)

Presenter: Francesco Tudisco (University of Edinburgh)

Abstract: Deep learning-based methods have shown remarkable effectiveness in solving PDEs, largely due to their ability to enable fast simulations once trained. However, despite the availability of high-performance computing infrastructure, many critical applications remain constrained by the substantial computational costs associated with generating large-scale, high-quality datasets and training models. In this work, inspired by studies on the structure of Green's functions for elliptic PDEs, we introduce Neural-HSS, a parameter-efficient architecture built upon the Hierarchical Semi-Separable (HSS) matrix structure that is provably data-efficient for a broad class of PDEs. We theoretically analyze the proposed architecture, proving that it satisfies exactness properties even in very low-data regimes. We also investigate its connections with other architectural primitives, such as the Fourier neural operator layer and convolutional layers. We experimentally validate the data efficiency of Neural-HSS on the three-dimensional Poisson equation over a grid of two million points, demonstrating its superior ability to learn from data generated by elliptic PDEs in the low-data regime while outperforming baseline methods. Finally, we demonstrate its capability to learn from data arising from a broad class of PDEs in diverse domains, including electromagnetism, fluid dynamics, and biology.

Bayesian Heston model estimation with LLM-guided evolutionary search

Track: World Class Research Track

Authors: Sean Yalcin (Queen's University Belfast), Pantelis Sopasakis (Queen's University Belfast)

Presenter: Sean Yalcin (Queen's University Belfast)

Abstract: This paper presents a Bayesian Heston-model estimation framework that jointly infers latent volatility paths and model parameters from asset prices. An LLM-guided evolutionary search, implemented with OpenEvolve, tunes priors and model parameters using a composite score that we propose, combining posterior predictive fit, convergence of MCMC, sampling efficiency, and posterior geometry penalising elongated and multimodal posteriors. Experiments on synthetic data and FTSE All-World Index prices show that the evolved models achieve well-calibrated posteriors and improved predictive performance.

Prediction of the infecting organism in peritoneal dialysis patients with acute peritonitis using interpretable Tsetlin Machines

Track: World Class Research Track

Authors: Anatoliy Gorbenko (Leeds Beckett University), Olga Tarasyuk (Newcastle University), Matthias Eberl (Cardiff University), Nicholas Topley (Cardiff University), Jingjing Zhang (Newcastle University), Rishad Shafik (Newcastle University), Alex Yakovlev (Newcastle University)

Presenter: Anatoliy Gorbenko (Leeds Beckett University)

Abstract: The analysis of complex biomedical datasets is becoming central to understanding disease mech-anisms, aiding risk stratification and guiding patient management. However, the utility of computational methods is often constrained by their lack of interpretability, which is particularly relevant in clinically critical areas where rapid initiation of targeted therapies is key. To define diagnostically relevant immune signatures in peritoneal dialysis patients presenting with acute peritonitis, we analysed a comprehensive array of cellular and soluble parameters in cloudy peritoneal effluents. Utilising Tsetlin Machines, a novel logic-based machine learning approach, we identified pathogen-specific immune fingerprints for different bacterial groups, each characterised by unique biomarker combinations. Unlike traditional ‘black box’ machine learning models, Tsetlin Machines identified clear, logical rules in the dataset that pointed towards distinctly nuanced immune responses to different types of bacterial infection. Importantly, these immune signatures could be easily visualised to facilitate their interpretation, thereby allowing for rapid, accurate and transparent decision-making. This unique diagnostic capacity of Tsetlin Machines could help deliver early patient risk stratification and support informed treatment choices in advance of conventional microbiological culture results, thus guiding antibiotic stewardship and contributing to improved patient outcomes.

Time Without Death: Finitude, Social Order, and What Machines Lack

Track: Emerging Research

Authors: Canhui LIU (University College London)

Abstract: Machine collectives increasingly coordinate, reciprocate, and form shared conventions on their own, and it is tempting to conclude that they are societies in the sense that human populations are. We argue that this conflates two registers of social order and misses what the human register is for. Our thesis is that human sociality is the way a finite, natal, generational form of life organises its own finitude. That finitude has three faces. Members die, and their tacit knowledge dies with them. Newcomers are ignorant by default. Cohorts are replaced and must hand on what they cannot keep. Kinship, inheritance, teaching, tradition, and much of obligation and trust are the forms this organisation takes, and machine collectives reproduce only what remains once finitude is subtracted. We address the obvious objection that machines can be given death, with two distinctions. The first is autonomy versus heteronomy. A death that an external operator can reset, roll back, or evade by copying is not constitutive finitude, and the test is resettability. The second is representation versus binding force. Because humans learn finitude socially, from the deaths of others, its behavioural representation is learnable by machines, but its grip on motivation requires that the learner's own end be inescapable. We treat machine collectives as a model organism for social theory and connect three levels of analysis. At the micro level, a controlled experiment holds task, learning rule, population, and compute fixed and varies only whether death means loss. Cumulative, transmissible culture arises only under irreversible loss, and a population that copies itself forward is individually more capable yet culturally empty. At the macro level, we compute the same descent signatures on a reproducing machine population and on three real human genealogies analysed through one pipeline: the China Biographical Database, European nobility, and the Mathematics Genealogy Project. Copying over-transmits status relative to every human regime, including hereditary nobility, and fragments lineage like blood descent, whereas externalisation lowers transmission into the human range and connects lineage like intellectual descent. With a real frontier model, the end-game defection of a language-model population vanishes when the same game is stripped of its labels, which is the signature of recognition rather than a learned mechanism. We position the work against evolutionary reinforcement learning, open-endedness, self-improving agents, and cultural accumulation in reinforcement learning,

and we conclude that the diachronic register is not emergent in present AI but constructible.

Understanding User Trust and Expectations in Answerable Sociotechnical Systems

Track: World Class Research Track

Authors: Dilara Kekulluoglu (Sabanci University), Yuqi Niu (Shanghai Jiao Tong University), Michael Rovatsos (University of Edinburgh), Nadin Kokciyan (University of Edinburgh)

Presenter: Dilara Kekulluoglu (Sabanci University)

Abstract: Today, we see an increase in the number of systems where Artificial Intelligence and Autonomous Systems (AI/AS) are included in decision-making processes, even in high-stakes domains such as healthcare. It is often not clear how to take responsibility appropriately and provide restitution when significant harms happen, especially when organizations use opaque AI/AS to make decisions. We focus on the notion of responsibility as answerability, which demands that individuals have a duty to supply their "reasons and justifications" for an action, or respond to inquiries about why they acted in a particular manner. We investigate the trust towards organizations that use mediator agents, people's expectations from such agents, perceived benefits and risks of using this new intervention to enable answerability in organizations. Our findings indicate that providing explanations and offering remedies for harm through mediator agents does not affect the level of trust individuals have in organizations.

Machine Translation for Patients with Limited English Proficiency: A Feasibility Study with Simulated Consultations

Track: Emerging Research

Authors: Yuliia Antiufieieva (University of Southampton), Adriana Wilde (University of Southampton)

Presenter: Yuliia Antiufieieva (University of Southampton)

Abstract: A growing number of people in the UK face barriers to healthcare access due to limited English proficiency (LEP), compounded by declining investment in interpreter services. Machine translation (MT) offers a rapid, low-cost means of facilitating multilingual medical communication. We present a pilot study assessing the accuracy and safety of MT for transcribed verbal communication in simulated medical consultations; a scenario that has received comparatively little attention in the literature relative to written document translation. Using a dataset of simulated GP, A&E, and pharmacy consultations, we compare the output quality of Google Translate and Copilot (GPT-5) across English-Ukrainian and English-Russian translation pairs. Contrary to expectation, Ukrainian (a low-resource language) produced only marginally more errors than Russian. Our findings suggest that the clinical actionability of machine translation output might be more important than its surface-level accuracy. This motivates future work examining patient comprehension of machine-translated content.

Convergence of an actor critic gradient flow for entropy regularised MDPs in general spaces

Track: World Class Research Track

Authors: Denis Zorba (University of Edinburgh)

Abstract: We prove the stability and global convergence of a coupled actor-critic gradient flow for infinite-horizon and entropy-regularised Markov decision processes (MDPs) in continuous state and action space with linear function approximation under Q-function realisability. We consider a version of the actor critic gradient flow where the critic is updated using temporal difference (TD) learning while the policy is updated using a policy mirror descent method on a separate timescale. For general action spaces, the relative entropy regularizer is unbounded and thus it is not clear a priori that the actor-critc flow does not suffer from finite-time blow-up. Therefore we first demonstrate stability which in turn enables us obtain a convergence rate of the actor critic flow to the optimal regularised value function. The arguments presented show that timescale separation is crucial for stability and convergence in this setting.

Method Development for Efficient Monitoring of Seafloor Disturbance from Bottom Trawling Using Artificial Intelligence and Autonomous Surveying.

Track: Emerging Research

Authors: Mouri Hoque Nadia (University of Southampton), Blair Thornton (University of Southampton), Susan Gourvenec (University of Southampton), Jasmin Godbold (University of Southampton)

Abstract: This study aims to develop novel methodologies for the efficient monitoring of seafloor disturbances resulting from bottom trawling. Recognising the limitations of current approaches that rely on indirect or low-resolution data, this research will investigate the integration of artificial intelligence with autonomous underwater vehicles (AUVs) to enable high-resolution, in situ detection and characterisation of benthic impacts. The focus will be on designing and validating scalable onboard data acquisition and interpretation techniques, capable of operating across diverse marine environments. Rather than producing global disturbance estimates directly, the contribution lies in advancing methods that can improve the precision, efficiency, and autonomy of seafloor monitoring. These tools are expected to support future large-scale assessments and inform evidence-based approaches to sustainable fisheries management and marine conservation.

Social Educational Robotics and Learning Analytics: A Scoping Review of an Emerging Field

Track: World Class Research Track

Authors: Pablo Gonzalez-Oliveras (KTH Royal Institute of Technology), Olov Engwall (KTH Royal Institute of Technology), Adriana Wilde (University of Southampton)

Presenter: Adriana Wilde (University of Southampton)

Abstract: Social Educational Robotics and Learning Analytics (LA) are prominent fields in technology-enhanced learning, but their combined potential remains underexplored, despite methodological similarities. Increasingly, signs of joint interests have emerged, with a surge in publications mentioning both social robots and learning analytics in the last five years. We therefore conducted a scoping review to explore if a new research field is emerging. We identified 29 empirical studies that combine social robots and LA, but also found that few studies explicitly state that social educational robots and LA are used in combination. Several studies used social educational robots that adapted to the learners or the learning environment based on interaction data. This signifies that they are in fact employing the feedback cycle that is at the core of LA methodology, but as most of these studies update the learner model using post-session data (e.g., learner improvement or feedback), they are long-term studies with repeated interventions that are applying LA methodology inadvertently. There are also benefits for LA research to use social educational robots, since LA increasingly uses an array of equipment to collect multimodal data, and all studies in this review employ at least two input modalities (= 4.4). Social robots provide the possibility to collect this data non-intrusively with the robot itself, in addition to creating a pedagogically boosted interaction compared to traditional LA interventions (e.g., learning management systems). By raising researchers' awareness of how close the fields of social educational robotics and LA are, substantial synergy effects could therefore be gained.

Effects of Swarm Size Variability on Operator Workload

Track: World Class Research Track

Authors: William Hunt (University of Southampton), Aleksandra Landowska (University of Nottingham), Horia Maior (University of Nottingham), Sarvapali D. Ramchurn (University of Southampton), Mohammad Soorati (University of Southampton)

Abstract: Real-world deployments of human–swarm teams depend on balancing operator workload to leverage human strengths without inducing overload. A key challenge is that swarm size is often dynamic: robots may join or leave the mission due to failures or redeployment, causing abrupt workload fluctuations. Understanding how such changes affect human workload and performance is critical for robust human–swarm interaction design. This paper investigates how the magnitude and direction of changes in swarm size influence operator workload. Drawing on the concept of workload history, we test three hypotheses: (1) workload remains elevated following decreases in swarm size, (2) small increases are more manageable than large jumps, and (3) sufficiently large changes override these effects by inducing a cognitive reset. We conducted two studies (N = 34) using a monitoring task with simulated drone swarms of varying sizes. By varying the swarm size between episodes, we measured perceived workload relative to swarm size changes. Results show that objective performance is largely unaffected by small changes in swarm size, while subjective workload is sensitive to both change direction and magnitude. Small increases preserve lower workload, whereas small decreases leave workload elevated, indicating workload residue; large changes in either direction attenuate these effects, suggesting a reset response. These findings offer actionable guidance for managing swarm-size transitions to support operator workload in dynamic human–swarm systems.

PARIS: Pruning Algorithm via the Representer theorem for Imbalanced Scenarios

Track: Emerging Research

Authors: Enrico Camporeale (Queen Mary University of London)

Presenter: Enrico Camporeale (Queen Mary University of London)

Abstract: The challenge of imbalanced regression arises when standard Empirical Risk Minimization (ERM) biases models toward high-frequency regions of the data distribution, causing severe degradation on rare but high-impact ``tail'' events. Existing strategies such as loss re-weighting or synthetic over-sampling often introduce noise, distort the underlying distribution, or add substantial algorithmic complexity. We introduce PARIS (Pruning Algorithm via the Representer theorem for Imbalanced Scenarios), a principled framework that mitigates imbalance by optimizing the training set itself. PARIS leverages the representer theorem for neural networks to compute a closed-form representer deletion residual, which quantifies the exact change in validation loss caused by removing a single training point without retraining. Combined with an efficient Cholesky rank-one downdating scheme, PARIS performs fast, iterative pruning that eliminates uninformative or performance-degrading samples. We use a real-world space weather example, where PARIS reduces the training set by up to 75\% while preserving or improving overall RMSE, outperforming re-weighting, synthetic oversampling, and boosting baselines. Our results demonstrate that representer-guided dataset pruning is a powerful, interpretable, and computationally efficient approach to rare-event regression.

Machine Learning Decoding of SYNGAP1 Genotype and Fear-State Neural Dynamics

Track: Emerging Research

Authors: Lachin Soufizadeh Balaneji (School of Informatics, University of Edinburgh Edinburgh), Arno Onken (School of Informatics, University of Edinburgh), Thomas Watson (Centre for Discovery Brain Sciences, University of Edinburgh), Peter Kind (Centre for Discovery Brain Sciences, University of Edinburgh), Paul Rignanese (Centre for Discovery Brain Sciences, University of Edinburgh)

Presenter: Lachin Soufizadeh Balaneji (School of Informatics, University of Edinburgh Edinburgh)

Abstract: This study investigates whether multi-region EEG and local field potential recordings can distinguish SYNGAP1 genotype and fear-related behavioural state in a rat model of SYNGAP1 haploinsufficiency. Spectral, complexity, and connectivity features from the medial prefrontal cortex, amygdala, and olfactory bulb were evaluated using five machine-learning classifiers. XGBoost achieved the best genotype-classification performance, while SHAP analysis highlighted gamma-band measures in the amygdala and olfactory bulb as influential features. Freezing behaviour was also classified accurately within each genotype using Elastic Net logistic regression. These findings show the potential of interpretable machine learning for identifying genotype- and behaviour-related electrophysiological signatures.

Machine Learning Assisted Physics Guided Analysis and Design Space Exploration of p-Channel GaN FETs

Track: Emerging Research

Authors: Fatih Arda Zengin (Sabanci University), Zarak Bhat (The University of Edinburgh), Shashank Mishra (The University of Edinburgh), Themis Prodromakis (The University of Edinburgh)

Presenter: Fatih Arda Zengin (Sabanci University)

Abstract: This work introduces a framework for the design-space exploration of p-channel GaN FETs, leveraging established relations from the literature and physics-based device simulations. Initially, the de- vice parameters such as gate length (L), oxide thickness ($t$ox), and GaN channel thickness (tGaN) are extracted from previously re- ported experimental and simulation based device characterisation data. These parameters are then used to define physically realistic simulation ranges and generate an extended input-output dataset linking device geometry with key electrical performance metrics, in- cluding maximum drain current (IDS,max), threshold voltage (Vth), on-resistance (Ron), and transconductance (gm). The developed database will then undergo data processing and feature engineering to make it suitable for machine learning assisted analysis. Finally, this framework provides a basis for design-space exploration and future inverse design workflows, where desired electrical perfor- mance targets can be mapped to feasible p-channel GaN FET design parameter ranges

Multistream Modelling for Mental Health: Modelling Linguistic and Temporal Contexts with Mutual and Self-Excitation in Social Media

Track: World Class Research Track

Authors: Anthony Hills (Queen Mary University of London), Talia Tseriotou (Queen Mary University of London), Mahmud Elahi Akhter (Queen Mary University of London), Junyu Mao (University of Southampton), Iqra Ali (Queen Mary University of London), Xenia Miscouridou (University of Cyprus), Maria Liakata (Queen Mary University of London)

Presenter: Talia Tseriotou (Queen Mary University of London)

Abstract: We present MHRoBERT (Multistream HEAT over Recurrence over BERT), a hierarchical transformer architecture for longitudinal mental health monitoring that models self- and mutual excitation patterns in linguistic and temporal data across multivariate event streams relating to an individual's mental health. To supply the model with complementary perspectives on each post, we apply a Large Language Model (LLM) based annotation to extract three streams from social media posts: emotional states, personal life events, and mental health symptoms. A central finding is that multi-task learning with these automatically-generated stream labels provides substantial, consistent improvements across all model architectures evaluated. Multistream information further consistently benefits simpler models not explicitly designed to exploit it: LLM baselines incorporating stream annotations improve macro F1 by 12.6% over text-only prompting. These results have direct implications on Moments of Change detection: multistream auxiliary supervision yields consistent, substantial gains regardless of architecture, suggesting it is a simple and portable strategy that future systems can readily adopt with minimal architectural changes. MHRoBERT additionally produces interpretable learned parameters across streams, revealing temporal interaction patterns between mental health indicators.

Persistent Homology for the Quantification of Microstructures in HRTEM Images

Track: Emerging Research

Authors: Ben Rowlinson (University of Edinburgh), Alexandros D. Keros (University of Edinburgh), Themis Prodromakis (University of Edinburgh)

Abstract: Analysis of Transmission Electron Microscopy images provides a useful source of materials science research, but conventional methods are sensitive to imaging conditions and rely on manual processing or parameter-dependent techniques. This work proposes an unsupervised approach to crystal segmentation based on Persistent Homology, a topological data analysis technique that extracts atomic structural information at multiple scales by investigating image sub level sets. Topological signatures are extracted from an image and clustered to distinguish between crystalline and amorphous regions, with sub clusters further investigated through persistence density and image processing. Building on prior applications of Persistent Homology in electron microscopy, this work implements an application to crystal segmentation in High Resolution Transmission Electron Microscopy imaging, providing a structure-driven method for materials analysis

Exploring the Use of AI-based Chatbots to Help Support Armed Forces Veterans' Mental Health

Track: World Class Research Track

Authors: Deepak Sahoo (Swansea University), Juan Maestre (Swansea University), Maheen Ashraf (Swansea University)

Presenter: Maheen Ashraf (Swansea University)

Abstract: Engagement with mental health services is often limited for Armed Forces Veterans (AFVs) due to stigma and cultural misalignment within civilian healthcare. While digital tools are proposed as a solution, empirical research on veteran-led evaluation of AI is scarce. We conducted semi-structured interviews with 9 AFVs and 2 mental health service providers to examine perceptions of AI chatbots and identify design implications. Findings reveal that veterans value chatbots as "cultural bridges" and low-threshold gateways that facilitate early disclosure and clinical continuity rather than autonomous treatment. Key requirements for adoption include interactional "cultural legitimacy" through military-specific language, alongside robust data sovereignty and user-controlled sensing. This paper argues that effective veteran-centric AI must prioritize cultural situatedness and trust over mere automation to successfully support help-seeking.

Exploring children's visions for an AI-enabled education

Track: World Class Research Track

Authors: Yve Smiith (Northumbria University)

Presenter: Yve Smiith (Northumbria University)

Abstract: This work presents ongoing PhD research exploring how children aged 7–11 envision AI-enabled education when scaffolded by creative AI literacy materials and speculative design activities built around an 'AI Creatures' concept. The project takes a Research through Design approach, developing engaging learning experiences that support children to critically engage with AI and contribute their own visions for the future of education. By positioning children as active contributors rather than passive recipients of AI, the work allows exploration of how their perspectives can inform the development of more human-centred educational technologies and contribute to wider discussions around responsible AI in education. The presentation will outline the research approach, the current iteration of materials and workshop programme, early implementations, and implications for AI literacy and children's participation in AI discourse.

Stellar: Multi-Agent Online Path Discovery for Post-Disaster Stochastic Graphs

Track: Emerging Research

Authors: Umut Yunusoğlu (University of Edinburgh), Sude Tıraş (Sabancı University), Cansu Akdaş (Sabancı University), Caglar Tunc (Sabancı University), Mehmet Murat Fadıloğlu (Sabancı University)

Presenter: Umut Yunusoğlu (University of Edinburgh)

Abstract: This work studies online route discovery in post-disaster networks where road or communication links may be damaged with uncertain severity. We propose Stellar, a multi-agent heuristic routing framework in which agents use sampled disaster-aware maps to explore diverse routes and share discovered edge states in real time. Experiments on stochastic grid graphs show that heuristic diversity improves the best-arriving agent but can increase tail cost, while inter-agent information sharing reduces the cost of the worst-arriving agents. The results suggest that the main benefit of Stellar is not uniformly shorter paths for all agents, but reduced variability and improved robustness across the fleet.

pyHyperRI: a PyTorch library for AI-accelerated radio astronomical imaging

Track: Emerging Research

Authors: Chao Tang (EPCC, University of Edinburgh), Adrian Jackson (EPCC, University of Edinburgh)

Presenter: Chao Tang (EPCC, University of Edinburgh)

Abstract: Radio-interferometric imaging is an ill-posed inverse problem. Efficient algorithms and high-performance implementations are crucial for handling extreme-scale wideband observations from next-generation radio telescopes such as the Square Kilometre Array. We present pyHyperRI, a PyTorch library for developing, benchmarking, and deploying radio astronomical imaging algorithms on modern accelerators. The library provides modular PyTorch-based interfaces, enabling native GPU execution and seamless integration with deep learning models. The library also supports distributed hyperspectral reconstruction by partitioning spectral channels across GPUs. We evaluate pyHyperRI on simulated hyperspectral imaging problems, showing that it reproduces the reconstruction quality of the MATLAB implementation while achieving approximately ten-fold acceleration.

Towards a Data-Centric AI Framework for Design of p-GaN E-mode HEMTs

Track: Emerging Research

Authors: Zarak Bhat (The University of Edinburgh)

Abstract: In this work, we have developed a curated database of experimental p-GaN HEMT structures, that would serve as a valuable resource for identifying geometrical device parameters, including gate length, gate-drain spacing, p-GaN thickness, AlGaN barrier thickness, and other pertinent structural features. The key performance metrics specific to power applications are, threshold voltage (Vth), on-resistance (Ron), breakdown voltage (BV), and saturation current (IDsat). These parameters are often reported under different bias conditions, units, extraction meth- ods, device geometries, and substrate platforms. Direct use of such data in machine-learning models may therefore lead to misleading correlations and physically inconsistent conclusions. The resulting cleaned dataset provides a foundation for developing surrogate models that relate device geometry to key power-device metrics. The long-term objective is to enable efficient forward prediction, sensitivity analysis, and inverse design of p-GaN E-mode HEMTs. This literature-based data-driven approach provides a pathway toward data-efficient, physics-informed device optimization for next-generation GaN power devices.

EDA-Trace: An Execution-Aware Schema for EDA Optimization Trajectories

Track: Emerging Research

Authors: Shu Gu (University of Edinburgh), Cristian Sestito (University of Edinburgh), Matthew Youngman (University of Edinburgh), Yuqin Zhao (University of Edinburgh), Ramesh Krishnamurthy (University of Edinburgh), Ali Siddique (University of Edinburgh), Themis Prodromakis (University of Edinburgh), Danial Chitnis (University of Edinburgh)

Presenter: Shu Gu (University of Edinburgh)

Abstract: EDA-Trace is an execution-aware schema for recording EDA optimisation attempts as structured state-action-execution-outcome trajectories. Instead of preserving only final QoR summaries, it captures candidate actions, attempted interventions, tool execution events, validation outcomes, QoR and runtime deltas, and checkpoint lineage. We instantiate the schema in a prototype trace-lake export layer around FPGA backend optimisation runs. On a LogicNets checkpoint, the prototype exports transition and candidate-action records from single-flow, fixed-trajectory, and frontier-search runs, with the best observed trajectory improving Fmax by 26.3%. The resulting traces support retrieval, candidate comparison, failure modelling, and future trajectory learning for ML4EDA and agentic EDA workflows.

Through the Lens of Doubt: Robust and Efficient Uncertainty Estimation for Visual Place Recognition

Track: World Class Research Track

Authors: Emily Miller (University of Southampton), Shoaib Ehsan (University of Southampton), Michael Milford (Queensland University of Technology), Sarvapali D. Ramchurn (University of Southampton), Muhammad Burhan Hafez (University of Southampton)

Presenter: Emily Miller (University of Southampton)

Abstract: Visual Place Recognition (VPR) enables robots and autonomous vehicles to identify previously visited locations by matching current observations against a database of known places. However, VPR systems face significant challenges when deployed across varying visual environments, lighting conditions, seasonal changes, and viewpoints changes. Failure-critical VPR applica- tions, such as loop closure detection in simultaneous localization and mapping (SLAM) pipelines, require robust estimation of place matching uncertainty. We propose three training-free uncertainty metrics that estimate prediction confidence by analyzing inherent statistical patterns in similarity scores from any existing VPR method. Similarity Distribution (SD) quantifies match distinctive- ness by measuring score separation between candidates; Ratio Spread (RS) evaluates competitive ambiguity among top-scoring lo- cations; and Statistical Uncertainty (SU) is a combination of SD and RS that provides a unified metric that generalizes across datasets and VPR methods without requiring validation data to select the optimal metric. All three metrics operate without additional model training, architectural modifications, or computationally expensive geometric verification. Comprehensive evaluation across nine state- of-the-art VPR methods and six benchmark datasets confirms that our metrics excel at discriminating between correct and incorrect VPR matches, and consistently outperform existing approaches while maintaining negligible computational overhead, making it deployable for real-time robotic applications across varied environ- mental conditions with improved precision-recall performance.

Measuring what matters? Revealing gaps and layers in Responsible AI metrics for the public sector

Track: Emerging Research

Authors: Denis Newman-Griffis (The University of Sheffiled), Carine Roos (The University of Sheffiled), shauna concannon (Durham University), Alex Albert (Independet Researcher), Susan Oman (The University of Sheffiled)

Abstract: Artificial intelligence (AI) systems are increasingly deployed across local government and central civil services to optimise operational efficiency. Logics of public value demand policy evaluation, yet assessing the costs and benefits of AI relies on technical and efficiency measures that do little to capture public values or the complexity of public service delivery [1, 2]. The 2026 update to government evaluation standards (known as the Magenta book) promises to “mov(e) the needle from "measuring what happened" to "learning how to improve in real-time" [3]. Measuring what matters: Scoping dimensions of measurement for responsible AI in the public sector [4] was a short project that coincided with the launch of the new Magenta book. AI evaluation is “particularly challenging in the public sector at present,” with “a gap between measurement and meaning”. [5, 6] Furthermore, “all public sector workers are feeling the impacts of AI” [5] This paper presents findings from this research 1, with local and central government workers to capture how AI “capability, capacity and confidence are differentiated” as well as how it should work in practice [2, 7] and 2, from a scoping analysis of measurement tools and evaluation frameworks. Drawing on empirical data from multiple qualitative approaches to understanding public sector metrics we map a multi-layered measurement landscape that, whilst linked, operates in isolation, and in a way that confuses and obfuscates the relationship between metrics in theory and their capacity to help us understand utility. Datasets that inform this paper include: 1, a scoping review of public sector metrics case studies (n69); 2, minutes from two co-productive workshops with central and local government stakeholders; 3, evaluation documentation, including toolkits and frameworks that were used as elicitation tools in the workshops, and annotated by participants; 4, reflective statements from the five authors of this paper. We argue that multi-layered systemic gaps plague public sector AI governance. These are as follows: first, the distinct layers of metrication, from international indices to localised metrics; second (and relatedly), high-level ethical evaluation frameworks and practical on-the-ground use cases; third, “differential capability, capacity and confidence” across the workforce; finally, the gap between those working with AI, and those most affected: the

general public who are serviced by these workers. We demonstrate that these layers are uncoordinated, failing to communicate or inform one another, and yet the ramification “re-perform” [8] across these layers, often widening, rather than reducing gaps. Consequently, current reporting practices are narrow and methodologically fragile, heavily favouring easily quantifiable metrics such as technical accuracy and short-term operational time-savings. Crucial dimensions of public value—including long-term societal outcomes, continuous fairness and bias monitoring, algorithmic explainability, public trust, and formal human oversight metrics—are critically omitted. As is the urgency to work with stakeholders, especially citizens, to ensure that these metrics are people-centred in public sector delivery. Furthermore, this measurement fragmentation exacerbates systemic lifecycle vulnerabilities. Evaluation is frequently co-opted as a symbolic, "one-off" pre-deployment compliance exercise or utilised for institutional "AI boosterism," rather than serving as an iterative mechanism for long-term accountability. This paper analyzes how technology-driven solutionism, fragmented administrative accountability, and acute vendor data asymmetries prevent public bodies from establishing robust baseline measurements or decommissioning strategies. We argue for problem-oriented pre-procurement interventions, longitudinal lifecycle monitoring, and multi-dimensional value assessments that pair technical indicators with qualitative, participatory approaches that can account for the people-centred requirements of public sector delivery. Ultimately, this paper outlines the need to prioritise institutional literacy and regional collaborative frameworks necessary to realign algorithmic measurement with democratic accountability and public trust.

Exploring Video-to-Radar Simulation Framework for Functional Monitoring in Dementia Care

Track: Emerging Research

Authors: Keniel Peart (University of Southampton), Heba Awad (University of Southampton), Kainat Yasmeen (University of Southampton), Shelly Vishwakarma (University of Southampton)

Abstract: Monitoring functional change in dementia is difficult because changes in mobility, routine, restlessness, and daily behaviour often emerge gradually and may be missed by periodic clinical assessments. Radar sensing offers a privacy-preserving way to monitor movement without recording identifiable images, but collecting real radar data from older adults in care settings is ethically and practically challenging. Existing radar datasets often rely on scripted laboratory actors, which may not capture real-world frailty, hesitation, or uncontrolled motion. This work explores a video-to-radar simulation framework that estimates human motion from existing activity videos, embodies it within mesh-based models, and generates synthetic radar signatures.

"It's All About Profiting": A Two-Phase Study Exploring AI Features and Menstrual Representation in Menstrual Tracking Applications

Track: Emerging Research

Authors: Faye Doughty (Northumbria University), Elizabeth Sillence (Northumbria University), Daniel Harrison (Northumbria University), Michael Smith (Northumbria University)

Presenter: Faye Doughty (Northumbria University)

Abstract: Menstrual tracking apps (MTAs), technologies designed to support people in understanding and learning about their menstrual health, are increasingly integrating artificial intelligence (AI) features to improve predictions, generate personalized insights, and identify symptoms of potential health conditions. However, there are concerning biases embedded in MTAs, including inconsistent predictions for people with irregular cycles, and gendered design, which could exclude and misinterpret varied menstrual experiences. Limited research has explored how MTAs represent AI features, and menstrual cycle experiences. In this two phase study, we aim to explore AI features and representation of menstrual experiences in five MTAs available on the iOS App Store through a critical content analysis and a nine-week longitudinal qualitative study.

A physics-informed neural network for the elastic buckling load of perforated plates

Track: Emerging Research

Authors: Elenor Naraidoo (University of Oxford), Barbara Rossi (University of Oxford)

Presenter: Elenor Naraidoo (University of Oxford)

Abstract: Design standards recommend that numerical or experimental methods should be used when codified semi-empirical approaches can unreliably model the behaviour of geometrically advanced members. This research argues that an appropriately trained physics-informed neural network could compute structural behaviour with the same high accuracy as the finite element method, but with much superior computational ease and speed. This argument is demonstrated on a thin-walled perforated plate subject to buckling, for which preliminary results demonstrate good agreement with FE predictions, achieving a median error of 3.23% in the critical buckling load while accurately reproducing the corresponding buckling mode shapes after only 250 training epochs.

An Explainable Transformer for Pan-Cancer Classification and Biomarker Discovery

Track: Emerging Research

Authors: Keniel Peart (University of Southampton), Joanna Grundy (University of Southampton)

Presenter: Keniel Peart (University of Southampton)

Abstract: This work presents an explainable AI framework for pan-cancer transcriptomic analysis. The model classifies 33 cancer subtypes from TCGA RNA-seq gene expression profiles using a transformer-based architecture that combines one-dimensional convolutional feature extraction with spatial self-attention. Rather than focusing only on predictive accuracy, the work addresses a key challenge in cancer genomics: understanding which genes drive a model's subtype prediction and what biological mechanisms those genes may suggest. After classification, gradient-based saliency maps are used to assign gene-level importance scores for each cancer subtype. Highly salient genes are then analysed as candidate biomarkers, cross-referenced with known cancer marker resources, and tested for functional enrichment using MSigDB and Fisher's exact test. The model achieves approximately 93% test accuracy while recovering known subtype biomarkers and identifying enriched biological functions, including cancer-relevant pathway associations. The contribution is an exploratory, explainability-first workflow for using deep learning as a hypothesis-generation tool in precision oncology. The system is not presented as a clinical decision model, but as a research instrument for prioritising genes, cancer subtypes, and biological functions for follow-up validation.

The Six Elements of the Design Space

Track: Emerging Research

Authors: Tommaso Turchi (University of Pisa, Italy), Ben Wilson (Swansea University, UK), Matt Roach (Swansea University, UK), Alan Dix (Cardiff Metropolitan University & Swansea University, UK), Alessio Malizia (University of Pisa, Italy), Peter Daish (Swansea University, UK)

Presenter: Matt Roach (Swansea University, UK)

Abstract: Most AI decision systems still operate within narrow combination paradigms: humans validate, override, or defer to AI recommendations. Some systems improve decision-making performance, but there is a gap, rarely bridged, between the observed outcome and the ideal potential where human-AI combinations exceed what either could accomplish alone. Rather than proposing yet another algorithmic solution, we sought to map the design space itself: What elements do designers need to create genuinely synergistic human-AI decision systems? How do these elements interact across different domains, contexts and stakeholder groups? What has the community learned from attempts to bridge the synergy gap? Current approaches often treat human-AI combination as an engineering problem: How can we make AI outputs more interpretable? How do we calibrate user trust? While these questions matter, they miss a crucial insight: effective human-AI combination emerges from carefully orchestrating multiple interacting elements, not just better algorithms or interfaces

3D Pose Estimation and Classification for Sign Language Recognition

Track: Emerging Research

Authors: Arezo Sanayee (University of Southampton), Keniel Peart (University of Southampton), Shelly Vishwakarma (University of Southampton)

Abstract: This work trains a domain-specific 3D pose estimator on over 800,000 sign language frames to achieve accurate poses, including hand and finger shape. It demonstrates that SMPL-X rotation parameters achieving 87% cross-dataset sign classification accuracy, outperforming joint trajectory features by 26 percentage points. This is a finding not previously reported in sign language classification literature.

Navigating the landscape of LLM evaluation: Current approaches, limitations and the way forward

Track: Emerging Research

Authors: Jenny Chim (Queen Mary University of London), Mahmud Elahi Akhter (Queen Mary University of London), Guneet Singh Kohli (Queen Mary University of London), Yuxiang Zhou (Queen Mary University of London), Michael Schlichtkrull (Queen Mary University of London), Maria Liakata (Queen Mary University of London), Robert Procter (Warwick University), Tom Sorell (Warwick University), Joshua Kelsall (Warwick University), Aislinn Bergin (University of Nottingham), Maria Waheed (University of Nottingham)

Presenter: Guneet Singh Kohli (Queen Mary University of London)

Abstract: Large Language Models and agentic systems are increasingly embedded in real-world workflows, yet evaluation remains either benchmark-centric or dependent on LLM-as-a-judge focused on final-answer correctness. We provide a critical overview of the main evaluation paradigms for single agent LLMs, discussing how these extend to agentic workflows and the increased challenges involved in the latter. We discuss automatic metrics, static benchmarks, LLMas-a-judge and highlight their technical limitations. We propose that the way foreward in terms of technical evaluation should focus on transparent reference-free frameworks that allow examining not just the final answer but how this was derived. We also argue that evluation needs to be continuous and not only performed once prior to deployment. Moreover while understanding the technical quality of generated outputs is of paramount importance in providing evidence for the reliability of a system it is only part of the story when it comes to establishing whether a deployed system is fit for purpose in the wider context in which it operates. We argue for multi-faceted evaluation of technical aspects but emphasise that current practice should not only measure the satisfiability of these, comparing different systems but also understand how systems fit in a wider workflow

STARS TFT: An Integrated Desktop Tool for Thin-Film-Transistor Characterization, Yield Analysis, and Recipe Management

Track: Emerging Research

Authors: Saim Egemen Yucel (University of Edinburgh), Alin Panca (University of Edinburgh), Ben Rowlinson (University of Edinburgh), Lai Gun (University of Edinburgh), Themis Prodromakis (University of Edinburgh)

Abstract: Thin-film transistor (TFT) characterization in research labs is typi- cally run through scattered spreadsheets and manual calculations, which makes parameter extraction slow, error-prone, and hard to reuse. We present STARS TFT, an integrated desktop tool that unifies the full characterization chain around a single database: ingesting raw I–V measurements, extracting device parameters (threshold voltage, subthreshold swing, transconductance, and sat- uration mobility), classifying each device, and mapping yield and hotspots across a wafer. The tool processes real, noisy lab data, including multi-run measurement files, gate-leakage-dominated regions, and measurement noise, and resolves common extraction failures such as non-physical mobility singularities and incorrect subthreshold-swing estimates. Because every raw data point, ex- tracted parameter, quality label, and fabrication recipe is stored as linked, structured data, the tool also produces a labeled dataset suitable for training machine-learning models that connect process parameters to device performance and yield. We describe the sys- tem architecture, its handling of real measurement data, and the path toward ML-driven process optimization.

Analysing Electronic Properties of Point Defects using ML

Track: Emerging Research

Authors: Dominic Stead (The University of Edinburgh), Ben Rowlinson (The University of Edinburgh), Zarak Bhat (The University of Edinburgh), Themis Prodromakis (The University of Edinburgh)

Abstract: Defects within materials have gained significant interest due to the ability they have to affect the properties of the host material. Of great interest is the electronic properties due to the implications for defect engineering; the process of controlling defects within a material to modulate and tailor its physical properties for a de- sired purpose, which is crucial for industries such as the electronic devices sector. Here, we construct five different ML models that showed a predictive power for the defect formation energy, which was predicted for two interstitial structures (tetrahedral and hexag- onal), and substitution defect structures of Si. An XGBoost model was the best performing with an R2 value of 0.990, MAE of 0.968, and an RMSE of 1.317. This model can be later used to gain physi- cal insight about other properties related to the defect formation energy such as the charge transition levels (CTLs) of the defect structure

A Multi-Agent Tool for Verification Plan Generation and Coverage Analysis of Open Hardware Specifications

Track: Emerging Research

Authors: Reanna McEldowney (APRIL AI Hub, University of Edinburgh), Ramesh Krishnamurthy (APRIL AI Hub, University of Edinburgh), Yuqin Zhao (APRIL AI Hub, University of Edinburgh), Themis Prodromakis (APRIL AI Hub, University of Edinburgh), Danial Chitnis (APRIL AI Hub, University of Edinburgh)

Presenter: Reanna McEldowney (APRIL AI Hub, University of Edinburgh)

Abstract: This paper presents a multi-agent large language model (LLM)-based framework for automatically generating verification plans and coverage analysis reports from structured hardware requirements. The framework takes a refined requirements JSON file as input and produces a traceable verification plan containing requirement-linked test cases, verification constraints, execution steps, expected results, priorities, and estimated coverage metrics. In addition, an edge-case analysis stage combines a deterministic weak-language checker with an LLM-based analysis agent to identify requirements that may be difficult to verify because of ambiguous, vague, or conditional language. The proposed framework was evaluated using an I2C specification and a secondary I2C requirements dataset containing injected weak-language faults. Experimental results demonstrate high-quality verification plan generation, achieving a precision of 0.98 and a recall of 0.86 while maintaining traceability and output consistency. For edge-case detection, the framework achieved a precision of 1.00 and a recall of 0.80, demonstrating its effectiveness in identifying ambiguous and difficult-to-verify requirements. Future work will extend the framework to larger open hardware specifications, including AMBA AXI and RISC-V, and enhance the coverage analysis agent to automatically identify missing or incomplete verification coverage.

AI-Assisted Verification-Relevant Information Extraction and Quality Evaluation for Semiconductor Protocol Specifications

Track: Emerging Research

Authors: Themis Prodromakis (University of Edinburgh), Danial Chitnis (University of Edinburgh), Yuqin Zhao (University of Edinburgh), Ramesh Krishnamurthy (University of Edinburgh), Elsa Wakfi (University of Edinburgh)

Abstract: Semiconductor protocol specifications are complex natural-language documents containing functional requirements, protocol behaviours, timing constraints, and implementation details that must be care- fully analysed before verification can begin. Manual specification analysis is time-consuming, relies heavily on engineering expertise, and is prone to inconsistencies and overlooked requirements. This paper presents an AI-driven framework for extracting and evalu- ating verification-relevant information from semiconductor proto- col specifications. The proposed workflow first converts specifica- tion documents into structured representations before identifying protocol requirements, transaction behaviours, timing constraints, reset conditions, and corner cases while filtering non-essential de- scriptive content. A dedicated quality evaluation stage assesses the completeness, relevance, and traceability of the extracted informa- tion before producing a structured verification-ready dataset for downstream verification planning and coverage-gap analysis. By combining automated extraction with quality evaluation, the pro- posed framework reduces manual effort, improves consistency, and provides a scalable foundation for AI-assisted hardware verification

Conversational Visual Explanations: Towards Appropriate Reliance in Human-Centred Explainable AI

Track: Emerging Research

Authors: Alex Blandin (Swansea University), Matt Roach (Swansea University), Daniele Doneddu (Swansea University)

Presenter: Alex Blandin (Swansea University)

Abstract: Explanations routinely raise a person's tendency to accept a correct AI prediction, yet they rarely improve their ability to reject an incorrect one, and can even inflate perceived understanding (the Illusion of Explanatory Depth) and thereby encourage over-reliance at precisely the moments when scepticism and alertness are required. We present Conversational Visual Explanations (ConVEx), a human-centred explainable AI system that delivers Interactive, Causal, Contrastive, and Contextual (IC3) explanations as a structured, turn-based conversation rather than as a static artefact pushed at the user. ConVEx pairs visual attributions (SHAP and partial dependence) with counterfactual "what if?" queries and natural-language "rules of thumb", disclosed progressively through pre-coded question buttons that introduce deliberate friction instead of the misleading fluency with free-text chat. We describe a planned study in which participants act as analysts evaluating a glass-box model's rail-delay predictions, and we measure appropriate reliance directly, through the lens of diagnostic testing (sensitivity, specificity, and the positive likelihood ratio), rather than through self-report alone. We outline a cross-arm design that probes whether explanation depth improves the error-detection capacity of the human-AI pair, engendering appropriate reliance on the AI alongside appropriate self-reliance.

Addressing sociotechnical limitations of LLMs: A platform for requirement-aligned evaluation of AI systems

Track: Emerging Research

Authors: Iqra Ali (Queen Mary University of London), Maria Liakata (Queen Mary University of London), Mahmud Elahi Akhter (Queen Mary University of London), Robert Procter (Warwick University), Aislinn Bergin (University of Nottingham), Maria Waheed (University of Nottingham), Guneet Singh Kohli (Queen Mary University of London), Jenny Chim (Queen Mary University of London), Sebastian Löbbers (Queen Mary University of London), Mingwei Shi (Queen Mary University of London)

Presenter: Sebastian Löbbers (Queen Mary University of London)

Abstract: Large Language Models (LLMs) generate fluent, human-like text across languages and modalities, yet carry well-documented limitations like bias, privacy leakage, brittle reasoning, and a lack of explainability that raise serious concerns in high-stakes domains such as mental health, healthcare, and law. Before such systems can be responsibly deployed or commissioned, those who build, procure, and regulate them need evidence that they are safe, reliable, and fit for a specific real-world purpose. The AdSoLve project develops multi-faceted evaluation methods including novel criteria, metrics, and tasks as well as transparent reference free evaluation and provides them together in an evaluation platform. This can be used to assess the effectiveness and limitations of LLMs in practice and in real-world settings, thus helping support developers in improving model safety and providing evidence for policy makers and commissioners who must decide whether a system is fit for purpose.

VLM-Guided Visual Place Recognition for Planet-Scale Geo-Localization

Track: World Class Research Track

Authors: Sania Waheed (University of Southampton)

Presenter: Sania Waheed (University of Southampton)

Abstract: This paper presents a hybrid framework for scalable image-based geo-localization that combines the semantic reasoning capabilities of Vision-Language Models (VLMs) with retrieval-based Visual Place Recognition (VPR). Instead of relying solely on VLM-generated coordinate predictions or traditional retrieval methods, the proposed approach uses VLM predictions as geographic priors to constrain the retrieval search space before performing visual matching and geographic re-ranking. The framework is compatible with multiple state-of-the-art VPR methods and requires no task-specific training. Extensive evaluation on the IM2GPS, IM2GPS3k, and GWS15k benchmarks demonstrates consistent improvements over previous state-of-the-art methods, achieving particularly strong gains at street- and city-level localization while maintaining scalability to planet-scale reference databases. The work highlights how combining foundation models with retrieval-based computer vision methods can produce more accurate, interpretable, and efficient geo-localization systems for real-world applications such as robotics, navigation, and autonomous systems.

A Bayesian State Classification Framework for Detecting Tree Mortality in the South Amazon

Track: Emerging Research

Authors: Laura Baracaldo (Lancaster University)

Presenter: Laura Baracaldo (Lancaster University)

Abstract: Tracking localized forest disturbances in the Amazon using satellite time-series data is inherently confounded by severe atmospheric conditions and heavy background interference. Standard changepoint algorithms frequently overfit these transient environmental artifacts, highlighting the need for statistically rigorous methods capable of integrating ecological domain knowledge directly into the inference process. In this study, we propose a novel hierarchical Bayesian state classification framework to detect abrupt tree degradation. By modeling specific baseline and mortality trajectories; characterized by a two-breakpoint decline and recovery phase; our method effectively isolates true observed mortality events from baseline phenology and weather-induced noise. Applied to high-resolution PlanetScope imagery from the South Amazon, our Bayesian framework incorporates a dual-index fusion strategy (NDVI and EVI) that yields calibrated posterior probabilities, successfully filtering atmospheric artifacts and outperforming standard frequentist benchmarks.

Integrated Data-Driven Micro-Mobility Systems: Combining Interactive Dashboard Analytics with Multi-Criteria Geospatial Routing for Smart Urban Navigation

Track: World Class Research Track

Authors: Devanjan Bhattacharya (University of Edinburgh)

Presenter: Devanjan Bhattacharya (University of Edinburgh)

Abstract: Urban mobility systems have undergone rapid transformation in recent years due to the convergence of digital technologies, geospatial data, and artificial intelligence. In rapidly expanding cities, particularly in the Global South, the challenges of congestion, sustainability, and accessibility have intensified. At the same time, the COVID-19 pandemic significantly altered mobility patterns by disrupting public transport systems and increasing reliance on micro-mobility modes such as cycling and walking. In this context, computational tools—especially mobility dashboards and application-based routing systems—have emerged as critical instruments for analysing and managing urban transport systems. The mobility platform examined in this study represents a data-driven environment for exploring cycling patterns, accessibility, and route planning in Bhubaneswar, India. Parallel to such platforms, recent research has proposed algorithmic models for safe navigation under pandemic conditions, integrating infection-risk data into routing decisions. This paper synthesizes these two strands, namely, the interactive mobility dashboards for understanding spatial and behavioural patterns; and, the geospatial routing algorithms that incorporate safety and risk factors. The central argument is that combining these approaches allows for a more holistic understanding of urban mobility: not only how people move, but how they should move under constraints such as safety, sustainability, and accessibility.

A framework for Responsible Geospatial AI

Track: Emerging Research

Authors: Efstathios Margaritis (University of Southampton), Rachel Higgins (Ordnance Survey), Ayushi Goyal (Amazon Web Services), Adarsh Valoor (University of Southampton), Sarvapali D. Ramchurn (University of Southampton), Jadu Dash (University of Southampton)

Presenter: Efstathios Margaritis (University of Southampton)

Abstract: Artificial Intelligence is rapidly changing how geospatial data are produced, analysed and reused, from Earth Observation workflows to automated feature extraction and generative mapping. Yet existing Responsible AI frameworks do not fully address the distinctive risks created by location data. This paper argues that GeoAI requires an additional governance layer grounded in spatial exceptionalism: the idea that spatial data create risks through re-identifiability, spatial leakage, place-based collective harm, temporal persistence and spatial aggregation. Drawing on work by Southampton Geospatial with Ordnance Survey, Amazon Web Services and Responsible AI UK, we introduce a Responsible GeoAI framework and propose a GeoAI Transparency Passport to record where AI intervenes in geospatial workflows, what transformations occur and how downstream reuse should be governed. The paper invites the UKAIRS community to treat location as a first-class concern in AI assurance.

Pretraining EHR Foundation Models with Patient-Aware Sampling

Track: Emerging Research

Authors: Joshua Placidi (Imperial College London), Yuxuan Liu (Imperial College London), Jinpei Han (Imperial College London), Marek Rei (Imperial College London), Aldo Faisal (Imperial College London)

Abstract: This work studies how pretraining sequence construction affects autoregressive foundation models for electronic health records. Standard EHR pretraining typically concatenates patient trajectories into a global token stream and samples fixed-length windows, which can mix multiple patients within a window and give longer patient records more influence during optimization. We propose Patient Sampling, a patient-aware sequence-construction strategy that first samples a patient and then samples a window from that patient alone, allowing control over how training signal is distributed across patients. Evaluating on MIMIC-IV v2.2 and v3.1 with zero-shot clinical prediction tasks, we find that Patient Sampling improves macro AUROC and AUPRC over the standard Global Stream baseline. These results highlight sequence construction as an important and underexplored design choice for EHR foundation models.

IRIDiS: Interpretable AI for Rapid Infectious Disease Screening and Test Design

Track: Emerging Research

Authors: Anatoliy Gorbenko (Leeds Beckett University), Olga Tarasyuk (Newcastle University), Matthias Eberl (Cardiff University), Nicholas Topley (Cardiff University), Oleksandr Gordieiev (Leeds Beckett University), Rishad Shafik (Newcastle University), Alex Yakovlev (Newcastle University)

Presenter: Anatoliy Gorbenko (Leeds Beckett University)

Abstract: Rapid and accurate identification of infectious pathogens at the point of care remains one of the most pressing challenges in modern medicine. This work proposes IRIDiS – an emerging research programme on Interpretable AI for Rapid Infectious Disease Screening and Test Design that has the potential to revolutionize the future of clinical diagnostics, making infectious disease testing affordable, massively scalable and largely performed by patients at home, much as lateral flow test (LFT) kits transformed COVID-19 testing globally. By coupling semi-quantitative LFT readouts with the Tsetlin Machine (TM), a symbolic logic-based, interpretable learning algorithm, IRIDiS will delivers transparent, clinician-validated diagnostic decisions and data-driven LFT design optimisation.