Postgraduate Certificate in AI in Health and Social Care

Ethics and Governance of AI in Health and Social Care

Artificial Intelligence (AI) – A broad field of computer science focused on creating systems that can perform tasks normally requiring human intelligence. Related terms: Machine Learning, Deep Learning, Neural Networks. Explanation: AI encompasses rule-based expert systems, statistical models, and adaptive algorithms that can analyze data, recognize patterns, and make decisions. In health and social care, AI can support diagnosis, treatment planning, and resource allocation. Example: An AI-driven imaging tool that flags potential lung nodules on chest X-rays. Practical application: Automating triage in emergency departments to prioritize patients based on severity. Challenges: Bias in training data, transparency of decision-making, and integration with existing clinical workflows.

Algorithmic Bias – Systematic and unfair discrimination that arises when an algorithm produces prejudiced outcomes. Related terms: Fairness, Disparate Impact, Ethical AI. Explanation: Bias can stem from skewed datasets, flawed feature selection, or inappropriate model assumptions, leading to unequal treatment of patient groups. Example: A predictive model that underestimates risk for minority patients because their historical data are under-represented. Practical application: Auditing AI tools before deployment to detect and mitigate bias. Challenges: Identifying hidden biases, balancing fairness with model performance, and maintaining ongoing monitoring.

Algorithmic Transparency – The degree to which the inner workings of an AI system are open and understandable to stakeholders. Related terms: Explainability, Interpretability, Open Source. Explanation: Transparency involves documenting data sources, model architecture, and decision pathways, enabling clinicians and patients to trust and verify AI outputs. Example: Providing a visual heatmap that shows which regions of an MRI contributed most to a diagnosis. Practical application: Regulatory submissions that require a “model card” describing performance across demographic groups. Challenges: Trade-offs between proprietary technology protection and the need for openness, especially in commercial products.

Automation Bias – The tendency of human users to over-rely on automated decisions, even when they are incorrect. Related terms: Human-in-the-Loop, Decision Support, Cognitive Bias. Explanation: When clinicians place undue confidence in AI recommendations, errors may go uncorrected, compromising patient safety. Example: A radiologist dismisses a visible fracture because the AI report marked the image as normal. Practical application: Designing interfaces that require active confirmation from clinicians before finalizing AI-generated suggestions. Challenges: Training staff to maintain critical judgment and designing alerts that are salient but not overwhelming.

Clinical Decision Support (CDS) – Tools that provide clinicians with patient-specific assessments or recommendations to aid decision-making. Related terms: Electronic Health Record (EHR), Decision-Aid Algorithms, Point-of-Care Analytics. Explanation: AI-enhanced CDS can synthesize large data sets, suggest evidence-based interventions, and reduce variability in care. Example: An AI system that predicts sepsis risk

and prompts early antibiotic administration. Practical application: Embedding risk scores within the EHR dashboard to guide treatment pathways. Challenges: Alert fatigue, integration with legacy systems, and ensuring the CDS respects patient autonomy.

Data Governance – The policies, standards, and procedures that manage the acquisition, storage, use, and disposal of data. Related terms: Data Stewardship, Data Ethics, Compliance. Explanation: Robust governance ensures data quality, security, and lawful processing, which are essential for trustworthy AI in health. Example: A hospital establishing a data-access committee that reviews requests for patient datasets used in AI research. Practical application: Implementing role-based permissions that limit who can view or modify sensitive health records. Challenges: Balancing data sharing for innovation with privacy obligations under regulations such as GDPR or HIPAA.

Data Minimisation – The principle of collecting only the data necessary for a specific purpose. Related terms: Purpose Limitation, Privacy by Design, Anonymisation. Explanation: Reducing data volume lowers privacy risk and simplifies compliance, while still enabling effective AI model training. Example: Using only age, gender, and lab results to predict medication adherence, rather than full genomic profiles. Practical application: Designing AI pipelines that discard extraneous fields before model ingestion. Challenges: Determining the minimal dataset that still yields accurate predictions, especially for complex clinical tasks.

Data Provenance – Documentation of the origin, lineage, and transformation history of data used in AI systems. Related terms: Metadata, Audit Trail, Traceability. Explanation: Knowing where data came from and how it was processed helps assess reliability, detect errors, and satisfy regulatory scrutiny. Example: Recording that a dataset of ECG recordings was sourced from a specific cardiology clinic, cleaned using a defined pipeline, and version-controlled. Practical application: Generating provenance reports for AI model certification. Challenges: Maintaining comprehensive logs across multiple data custodians and ensuring that provenance information is kept up-to-date.

Data Sovereignty – The concept that data are subject to the laws and governance structures of the jurisdiction where they are stored. Related terms: Cross-Border Data Transfer, National Health Data Policies, Cloud Computing. Explanation: Health data may be restricted from leaving the country, impacting AI development that relies on large, multinational datasets. Example: A UK NHS trust must store patient data on servers located within the United Kingdom to comply with the Data Protection Act. Practical application: Deploying federated learning frameworks that keep raw data on local servers while sharing model updates. Challenges: Coordinating multi-national collaborations, reconciling differing legal regimes, and managing latency in distributed training.

Data Stewardship – The responsible management and oversight of data assets throughout their lifecycle. Related terms: Data Governance, Data Custodian, Data Quality. Explanation: Stewards ensure data integrity, accessibility, and compliance, acting as custodians for both technical and ethical aspects. Example: A data steward who validates the completeness of a cancer registry before it is used to train a survival-prediction model. Practical application: Establishing stewardship roles within AI project teams to oversee data handling procedures. Challenges: Allocating sufficient resources, maintaining expertise across clinical and technical domains, and aligning stewardship with research timelines.

De-identification – The process of removing or obfuscating personal identifiers from datasets to protect privacy. Related terms: Anonymisation, Pseudonymisation, Privacy-Enhancing Technologies. Explanation: De-identification reduces re-identification risk while preserving data utility for AI training. Techniques include masking, generalisation, and noise addition. Example: Replacing patient names with random alphanumeric codes and aggregating exact birth dates to age groups. Practical application: Sharing de-identified imaging data with external AI vendors for algorithm development. Challenges: Balancing data utility with privacy, handling rare disease data where de-identification may still allow identification, and complying with evolving legal standards.

Deep Learning – A subset of machine learning that uses multilayered neural networks to model complex patterns. Related terms: Artificial Neural Networks, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs). Explanation: Deep learning excels at processing unstructured data such as medical images, speech, and free-text notes. It requires large labelled datasets and substantial computational resources. Example: A CNN that classifies skin lesions as benign or malignant from dermoscopic photographs. Practical application: Automating pathology slide analysis to assist histopathologists. Challenges: Opacity of model decisions, high data demands, susceptibility to adversarial attacks, and the need for extensive validation before clinical use.

Ethical AI – The design, development, and deployment of AI systems that uphold moral principles such as beneficence, non-maleficence, autonomy, and justice. Related terms: Responsible AI, AI Ethics Frameworks, Human-Centred Design. Explanation: Ethical AI in health and social care ensures that technology serves patient welfare, respects rights, and reduces inequities. Example: An AI triage tool that incorporates fairness constraints to avoid disadvantaging older adults. Practical application: Conducting ethical impact assessments before launching an AI-enabled telehealth platform. Challenges: Translating abstract ethical values into concrete technical specifications, and reconciling competing stakeholder priorities.

Explainable AI (XAI) – Methods and techniques that make AI model outputs understandable to human users. Related terms: Interpretability, Model Explainability, Post-hoc Analysis. Explanation: XAI provides insights such as feature importance, counterfactual explanations, or visual saliency maps, enabling clinicians to assess the plausibility of AI recommendations. Example: A SHAP (SHapley Additive exPlanations) plot that shows which lab values contributed most to a diabetes risk score. Practical application: Integrating explanation modules into decision-support dashboards to foster trust. Challenges: Maintaining explanation fidelity while preserving model performance, and avoiding information overload for end-users.

Federated Learning – A machine-learning approach where models are trained across multiple decentralized devices or servers holding local data samples, without exchanging the raw data. Related terms: Distributed Training, Privacy-Preserving Machine Learning, Edge AI. Explanation: Federated learning enables collaborative AI development across institutions while respecting data sovereignty and privacy constraints. Example: Hospitals jointly training a COVID-19 outcome predictor by sharing model weight updates instead of patient records. Practical application: Building a national AI model for early detection of dementia using data stored within each care provider's firewall. Challenges: Handling heterogeneous data quality, ensuring convergence of the global model, and safeguarding against malicious updates.

Human-in-the-Loop (HITL) – An interaction paradigm where humans actively supervise, intervene, or

validate AI system outputs. Related terms: Decision Support, Oversight, Collaborative Intelligence. Explanation: HITL safeguards against autonomous errors, preserves professional accountability, and leverages complementary strengths of humans and machines. Example: A radiology AI flag that requires a radiographer's sign-off before the report is finalized. Practical application: Implementing a verification step where a nurse reviews AI-generated medication dosage recommendations. Challenges: Designing seamless workflows that do not impede efficiency, and preventing over-reliance on AI.

Informed Consent for AI-Enabled Care – The process of providing patients with clear information about AI involvement in their treatment, enabling voluntary agreement. Related terms: Patient Autonomy, Transparency, Data Use Agreements. Explanation: Consent must cover data collection, algorithmic processing, potential risks, and the right to opt-out, aligning with ethical standards and legal mandates. Example: A consent form that explains a predictive analytics tool will be used to assess risk of readmission and that patients can decline its use. Practical application: Embedding consent options within patient portals for AI-driven remote monitoring programs. Challenges: Communicating technical concepts in lay language, managing consent revocation, and ensuring consistent documentation across care pathways.

Institutional Review Board (IRB) / Research Ethics Committee (REC) – Bodies that review and approve research involving human participants, including AI studies. Related terms: Ethical Review, Regulatory Oversight, Human Subjects Protection. Explanation: IRBs assess risk-benefit ratios, data protection measures, and participant rights, ensuring that AI research adheres to ethical norms. Example: An IRB evaluating a trial where an AI algorithm predicts postoperative complications and informs surgical planning. Practical application: Submitting a detailed protocol that outlines data handling, algorithm validation, and patient safeguarding procedures. Challenges: Keeping review processes up-to-date with rapidly evolving AI technologies, and balancing innovation with participant protection.

Interoperability – The ability of different information systems, devices, and applications to exchange, interpret, and use data cohesively. Related terms: FHIR (Fast Healthcare Interoperability Resources), Standardised APIs, Semantic Compatibility. Explanation: Interoperability enables AI tools to access diverse health datasets, integrate with EHRs, and share insights across care settings. Example: An AI-driven medication reconciliation engine that pulls data from pharmacy, lab, and primary-care systems using FHIR standards. Practical application: Deploying a cloud-based AI service that can be called via standardised RESTful APIs from any certified EHR. Challenges: Harmonising disparate data models, handling legacy systems, and ensuring security during data exchange.

Justifiable AI – An AI system whose design, deployment, and outcomes can be ethically and legally defended. Related terms: Accountability, Compliance, Risk Management. Explanation: Justifiability requires documenting the rationale for algorithmic choices, demonstrating alignment with clinical guidelines, and providing evidence of benefit. Example: Publishing a white-paper that details the validation study, performance metrics, and mitigation strategies for an AI sepsis alert. Practical application: Including a justification dossier in procurement contracts for AI vendors. Challenges: Anticipating future regulatory changes, maintaining documentation over the system's lifecycle, and addressing unforeseen harms.

Legal Liability – The legal responsibility for damages caused by AI-mediated decisions or actions. Related terms: Negligence, Product Liability, Professional Responsibility. Explanation: Determining who is liable—

clinician, institution, or AI vendor—depends on factors such as control, oversight, and contractual arrangements. Example: A lawsuit alleging that an AI-driven diagnostic error led to delayed cancer treatment. Practical application: Drafting indemnity clauses that clarify responsibilities between health providers and AI developers. Challenges: Ambiguities in existing law, the evolving nature of AI capabilities, and cross-jurisdictional issues.

Model Drift – The gradual degradation of an AI model's performance over time due to changes in underlying data distributions. Related terms: Concept Drift, Performance Monitoring, Model Retraining. Explanation: In health care, shifts may arise from new clinical protocols, population health changes, or technology updates, necessitating ongoing vigilance. Example: An AI model trained on pre-COVID-19 data that underestimates respiratory failure risk after the pandemic altered patient profiles. Practical application: Implementing automated dashboards that flag significant drops in predictive accuracy. Challenges: Detecting subtle drift, allocating resources for timely model updates, and ensuring updated models retain regulatory compliance.

Model Governance – The set of policies, procedures, and controls that oversee the development, deployment, and maintenance of AI models. Related terms: Model Lifecycle Management, Version Control, Compliance Audits. Explanation: Governance frameworks define responsibilities, documentation standards, risk assessments, and post-deployment monitoring to assure safe AI use. Example: A hospital establishing a Model Governance Board that reviews all AI models before clinical integration. Practical application: Maintaining a registry of model versions, performance metrics, and approved use cases. Challenges: Coordinating multidisciplinary stakeholders, keeping governance lightweight enough to avoid stifling innovation, and integrating with existing quality-improvement processes.

Patient-Centred AI – AI solutions designed with the needs, preferences, and rights of patients as primary considerations. Related terms: Human-Centered Design, Shared Decision-Making, User Experience (UX). Explanation: Patient-centred AI involves co-creation with patients, transparent communication about AI roles, and mechanisms for feedback and opt-out. Example: A mobile health app that uses AI to tailor activity recommendations, while clearly showing users how their data influence suggestions. Practical application: Conducting usability testing with diverse patient groups before rollout. Challenges: Addressing digital literacy gaps, ensuring accessibility for vulnerable populations, and preventing algorithmic paternalism.

Privacy-Preserving Machine Learning (PPML) – Techniques that enable model training while protecting individual privacy. Related terms: Homomorphic Encryption, Secure Multiparty Computation, Differential Privacy. Explanation: PPML methods limit exposure of raw data, often by encrypting inputs, adding statistical noise, or performing computations on encrypted data. Example: Applying differential privacy to a predictive model for hospital readmission, guaranteeing that the inclusion of any single patient does not significantly affect outputs. Practical application: Deploying a PPML pipeline that allows research institutions to collaboratively improve an AI model without sharing patient records. Challenges: Balancing privacy guarantees with model accuracy, computational overhead, and the complexity of implementation.

Regulatory Compliance – Adherence to laws, regulations, and standards governing AI use in health and social care. Related terms: GDPR, HIPAA, Medical Device Regulation (MDR). Explanation: Compliance requires systematic assessment of data protection, safety, efficacy, and post-market surveillance obligations.

Example: Submitting a conformity assessment dossier to the European Medicines Agency for an AI-based diagnostic device. Practical application: Conducting regular gap analyses to ensure ongoing alignment with evolving regulatory expectations. Challenges: Navigating fragmented international regulations, interpreting ambiguous guidance, and allocating resources for continuous compliance.

Responsible AI – A comprehensive approach that incorporates ethical, legal, technical, and societal considerations throughout the AI lifecycle. Related terms: Ethical AI, AI Governance, Sustainable AI. Explanation: Responsible AI frameworks provide checklists for fairness, transparency, accountability, robustness, and societal impact, guiding developers and users alike. Example: A health-system adopting the WHO's "Ethics and Governance of AI for Health" guidance as a baseline for all AI projects. Practical application: Embedding responsible-AI checkpoints into project management tools such as stage-gate reviews. Challenges: Translating high-level principles into concrete actions, measuring compliance, and fostering a culture that values responsible practices.

Risk Assessment – Systematic identification, analysis, and mitigation of potential harms associated with AI deployment. Related terms: Threat Modeling, Impact Analysis, Mitigation Strategies. Explanation: In health care, risk assessment examines clinical safety, data privacy, operational disruption, and reputational consequences. Example: Conducting a Failure Modes and Effects Analysis (FMEA) on an AI-driven medication dosing system. Practical application: Developing a risk-register that tracks identified hazards, likelihood, severity, and mitigation actions. Challenges: Anticipating rare but high-impact failure modes, updating assessments as models evolve, and integrating risk management with clinical governance.

Safety-Critical AI – AI applications where failures could result in serious harm or loss of life, such as diagnostic or therapeutic decision tools. Related terms: High-Integrity Systems, Medical Device Software, Reliability Engineering. Explanation: Safety-critical AI demands rigorous validation, redundancy, and fail-safe mechanisms to meet stringent standards. Example: An AI algorithm that autonomously adjusts insulin pump delivery based on continuous glucose monitoring. Practical application: Implementing dual-mode operation where the AI output is cross-checked by a clinician before acting on the pump. Challenges: Achieving regulatory approval, ensuring real-time performance, and providing robust evidence of safety under diverse clinical conditions.

Social Determinants of Health (SDOH) in AI – Factors such as income, education, housing, and environment that influence health outcomes and must be considered in AI modeling. Related terms: Equity, Bias Mitigation, Population Health. Explanation: Ignoring SDOH can lead to models that misrepresent risk for disadvantaged groups; incorporating these variables promotes fairness and accuracy. Example: Including zip-code–derived socioeconomic indicators when predicting chronic disease progression. Practical application: Designing AI dashboards that display disparities in predicted outcomes across demographic strata. Challenges: Accessing reliable SDOH data, handling privacy concerns, and avoiding reinforcement of stereotypes.

Stakeholder Engagement – Involving all relevant parties—patients, clinicians, technologists, regulators, and the public—in AI development and oversight. Related terms: Co-Design, Public Consultation, Governance Boards. Explanation: Engaged stakeholders provide diverse perspectives, improve trust, and help identify unintended consequences early. Example: Holding workshops with caregiver groups to gather feedback on

an AI-enabled home-monitoring system. Practical application: Forming an advisory council that meets quarterly to review AI project progress and ethical considerations. Challenges: Managing conflicting interests, ensuring representation of marginalized voices, and maintaining ongoing dialogue beyond initial phases.

Transparency Reporting – Publication of detailed information about AI system design, data sources, performance, and governance. Related terms: Model Cards, Data Sheets, Open Documentation. Explanation: Transparency reports enable scrutiny by regulators, clinicians, and the public, fostering accountability and trust. Example: A health-tech firm releases a model card describing accuracy, calibration, and demographic performance for its pneumonia detection AI. Practical application: Including transparency sections in procurement documents that require vendors to disclose model provenance. Challenges: Balancing commercial confidentiality with openness, avoiding information overload, and updating reports as models evolve.

Trustworthy AI – AI that is lawful, ethical, and robust, meeting criteria such as fairness, accountability, transparency, and resilience. Related terms: Ethical AI, Responsible AI, AI Assurance. Explanation: Trustworthiness is built through rigorous validation, clear communication, and mechanisms for redress when harms occur. Example: An AI system that provides a confidence score alongside each prediction, allowing clinicians to gauge reliability. Practical application: Conducting third-party audits that certify AI tools as trustworthy before they are adopted in clinical pathways. Challenges: Defining measurable trust metrics, maintaining trust over time as technology and contexts change, and dealing with public perception.

Unintended Consequences – Outcomes that were not anticipated during AI design, which may be beneficial or harmful. Related terms: Side Effects, Emergent Behavior, Impact Assessment. Explanation: In health care, unintended consequences can include workflow disruptions, changes in patient-provider relationships, or new forms of bias. Example: An AI scheduling system that inadvertently reduces appointment availability for certain geographic regions. Practical application: Performing scenario-based testing to uncover potential indirect effects before full deployment. Challenges: Predicting complex system interactions, allocating resources for post-implementation monitoring, and responding swiftly to negative impacts.

Validation (Technical and Clinical) – The process of confirming that an AI model performs as intended, both in algorithmic terms and within real-world clinical settings. Related terms: Verification, Performance Evaluation, External Validation. Explanation: Technical validation checks data integrity, algorithmic correctness, and statistical robustness; clinical validation assesses safety, efficacy, and usefulness in patient care. Example: Conducting a multi-centre study that compares AI-predicted stroke outcomes against actual patient trajectories. Practical application: Publishing validation results in peer-reviewed journals and submitting them to regulatory bodies. Challenges: Securing diverse datasets for external validation, avoiding over-fitting, and ensuring reproducibility across sites.

Virtual Care AI – AI technologies that support remote health services, such as telemedicine triage, symptom checkers, and virtual monitoring. Related terms: Telehealth, Remote Patient Monitoring, Digital Therapeutics. Explanation: Virtual care AI can extend access, reduce travel burdens, and enable continuous health surveillance, but must address data security and digital equity. Example: An AI chatbot that screens for mental-health concerns and routes users to appropriate clinicians. Practical application: Integrating

AI-driven risk alerts into video-consultation platforms to prompt clinicians when vital signs indicate deterioration. Challenges: Ensuring reliable internet connectivity, maintaining patient privacy in home settings, and validating AI performance without in-person examinations.

Whistleblower Protection in AI Governance – Safeguards for individuals who expose wrongdoing or unsafe practices related to AI systems. Related terms: Ethical Reporting, Organisational Culture, Compliance Hotline. Explanation: Protecting whistleblowers encourages early detection of ethical breaches, data misuse, or safety lapses in AI deployments. Example: A data scientist reports that an AI model used for resource allocation is systematically disadvantaging a specific patient group. Practical application: Establishing confidential reporting channels and clear policies that prohibit retaliation. Challenges: Cultivating a culture of openness, ensuring reports are investigated promptly, and balancing confidentiality with the need for transparent remediation.