INNOVATION & RESEARCH CAUCUS

SHAPING RESPONSIBLE INNOVATION THROUGH AI

Evidence and Recommendations for UKRI

IRC Report No: 090

REPORT PREPARED BY

Bowei Chen
University of Glasgow

Nuran Acur
University of Glasgow

Shuying Liu
University of Glasgow

Jack Leahy
Department of Science, Innovation and Technology

Delivered with ESRC and Innovate UK

Authors

- Professor Bowei Chen - Adam Smith Business School, University of Glasgow
- Professor Nuran Acur - Adam Smith Business School, University of Glasgow
- Shuying Liu - Adam Smith Business School, University of Glasgow
- Jack Leahy - Department of Science, Innovation and Technology

This document relates to IRC Project FFThree005: Shaping Responsible Innovation through AI: Evidence from UKRI-Funded Social Science Projects

Acknowledgements

This work was supported by Economic and Social Research Council (ESRC) grant ES/X010759/1 to the Innovation and Research Caucus (IRC).

We are very grateful to the project sponsors at UK Research & Innovation (UKRI) for their input into this research. The interpretations and opinions within this report are those of the authors and may not reflect the policy positions of UKRI or its constituent councils. We would also like to acknowledge and appreciate the efforts of the IRC Project Administration Team for their support in preparing this report.

About the Innovation and Research Caucus

The Innovation and Research Caucus supports the use of robust evidence and insights in UKRI's strategies and investments, as well as undertaking a co-produced programme of research. Our members are leading academics from across the social sciences, other disciplines and sectors, who are engaged in different aspects of innovation and research systems. We connect academic experts, UKRI, IUK and the (ESRC), by providing research insights to inform policy and practice. Professor Tim Vorley and Professor Stephen Roper are Co-Directors. The IRC is funded by UKRI via the ESRC and IUK, grant number ES/X010759/1. The support of the funders is acknowledged. The views expressed in this piece are those of the authors and do not necessarily represent those of the funders.

Contact

You are also welcome to email us if you have any questions about this report or the work of the IRC generally: info@ircaucus.ac.uk

Cite as: Chen, B. et al. 2026. Shaping Responsible Innovation Through AI. Evidence and Recommendations for UKRI. Oxford, UK: Innovation and Research Caucus

CONTENTS

EXECUTIVE SUMMARY

Artificial intelligence (AI) is becoming an increasingly important capability in publicly funded social science research, creating new opportunities while raising questions about transparency, governance, accountability, fairness, privacy and responsible innovation. This report analyses 3,287 AI-related UKRI-funded social science projects from the Gateway to Research Plus (GtR+) dataset, examining how AI methods have evolved, how they vary across disciplines, and how responsible AI concerns are reflected across the research portfolio.

Key takeaways:

- **AI adoption is accelerating and becoming more layered.** Statistical, computational and symbolic approaches remain important, while machine learning, deep learning and natural language processing have expanded rapidly, followed more recently by foundation and generative AI.

- **AI use varies across disciplines and research pathways.** Different methods reflect different research questions and data, and contribute to diverse outputs and impacts, including publications, datasets, software, collaborations and policy engagement.

- **Responsible AI is emerging but remains unevenly embedded.** Among the 3,287 projects analysed, 623 explicitly reference issues such as privacy and security, governance, transparency, trust, fairness, ethics, inclusion and safety.

- **Methodological innovation and responsible innovation are not yet fully aligned.** The findings highlight an opportunity for UKRI to embed responsible innovation more systematically as AI capabilities develop across its research portfolio.

- **A lifecycle approach can support more responsible AI-enabled research.** The report proposes a UKRI responsible innovation framework based on six principles—anticipation, reflexivity, inclusivity, responsiveness, transparency and equitability—to support evaluation from proposal development through implementation, impact assessment and organisational learning.

Overall, the findings position AI as a strategic research capability and provide UKRI with an evidence base for monitoring its evolution, strengthening responsible AI governance and informing future funding priorities.

1. Introduction

Artificial intelligence (AI) is rapidly transforming the way research is conceived, designed and conducted across the social sciences. AI methods such as machine learning, natural language processing, large language models, predictive modelling, simulation, computer vision and generative AI are no longer confined to computer science but are increasingly embedded within research workflows, supporting literature synthesis, data collection, qualitative and quantitative analysis, forecasting, decision support and knowledge generation. These technologies are expanding the methodological toolkit available to researchers, enabling new forms of interdisciplinary inquiry, facilitating the analysis of increasingly large and complex datasets, and opening entirely new avenues for addressing societal challenges.

For research funders such as UK Research and Innovation (UKRI), understanding how AI is being adopted within publicly funded research has become strategically important. AI is not simply another research tool; it is reshaping research methods, influencing collaboration patterns, creating new opportunities for scientific discovery, and changing how knowledge is produced and translated into societal impact. As AI capabilities continue to evolve, evidence is needed to understand which AI methods are being adopted, where they are being applied, how rapidly they are diffusing across disciplines, and whether their use reflects the principles of responsible and trustworthy research. Such evidence enables funders to identify emerging research capabilities, anticipate future methodological trends, support capacity building, inform investment strategies, and develop governance approaches that encourage innovation while safeguarding research integrity and public trust. Beyond UKRI, this knowledge is equally valuable for universities, researchers and policymakers seeking to understand how AI is transforming the research ecosystem and creating new opportunities for scientific discovery and societal impact.

At the same time, the increasing use of AI introduces important governance challenges. Questions about transparency, accountability, fairness, explainability, privacy, intellectual property, bias and societal impact have become central to discussions about the future of AI-enabled research. These challenges are particularly significant within the social sciences, where research frequently involves sensitive data, vulnerable populations and complex societal issues requiring careful ethical consideration. Consequently, understanding how AI is used is as important as identifying which AI methods are employed, since different methodological choices create different opportunities, risks and governance requirements.

Within the UK research and innovation system, these challenges are increasingly addressed through the framework of responsible innovation. Rather than focusing solely on research outputs, responsible innovation emphasises how research and innovation are conceived, designed, conducted and governed. It encourages researchers and institutions to anticipate potential societal implications, reflect critically on underlying assumptions and values, engage a diverse range of stakeholders, and respond adaptively to emerging knowledge and public concerns (Owen, Macnaghten and Stilgoe, 2012; Stilgoe, Owen and Macnaghten, 2013).

These principles have become central to UKRI and the Economic and Social Research Council (ESRC) in promoting research that is scientifically robust, socially responsible and responsive to public needs.

As AI technologies have become embedded within research processes, the broader principles of responsible innovation have increasingly been translated into discussions of responsible AI. Although definitions vary across the literature, responsible AI generally refers to the design, development, deployment and use of AI systems in ways that are transparent, fair, accountable, trustworthy, privacy-preserving and socially beneficial (Carbajal-Pina & Acur, 2025). In this report, responsible AI is understood not as a separate governance framework but as one practical mechanism through which the broader principles of responsible innovation are operationalised within AI-enabled research. Accordingly, this study examines not only the diffusion of AI methods but also the extent to which responsible innovation principles are reflected in project design, methodological choices and research reporting.

Despite the strategic importance of AI for the future of research, there remains limited empirical evidence on how AI methods are being adopted across publicly funded social science projects. Much of the existing literature focuses on normative governance frameworks, ethical principles or technical developments, while comparatively little is known about how AI is actually being used within funded research portfolios, how transparently researchers report their AI methods, whether adoption differs across disciplines, how AI-enabled research has evolved over time, or the extent to which responsible innovation principles are evident within research practice. Without such evidence, policymakers and research funders have only a partial understanding of how AI is transforming research practice and whether responsible innovation is becoming embedded within everyday research activities rather than remaining primarily a policy aspiration.

This report addresses these gaps by systematically analysing AI adoption across UKRI-funded social science research using the Gateway to Research Plus (GtR+) database. By examining project metadata, abstracts, technical summaries and impact statements, the study provides a large-scale empirical assessment of how AI methods are being integrated into publicly funded social science research. Beyond identifying AI-enabled projects, the analysis investigates the methodological choices researchers make, the transparency with which AI methods are reported, the emergence of responsible AI themes, and the relationship between methodological innovation and responsible innovation practices.

Specifically, the study is guided by five interrelated research questions that together provide a comprehensive understanding of both AI adoption and the operationalisation of responsible innovation within the UK research and innovation system.

RQ1. How transparently do project abstracts disclose the AI methods they employ?

Transparent reporting is fundamental for research integrity, reproducibility and accountability. Understanding how clearly researchers describe their AI methods enables UKRI to assess whether funded projects provide sufficient methodological information for evaluation, knowledge sharing and responsible governance.

RQ2. Which AI methods are most prevalent across the funded research portfolio?

Identifying the dominant AI approaches reveals the evolving methodological landscape of AI-enabled social science research, highlights emerging research capabilities, identifies new methodological trends and provides evidence for future investment, skills development and research infrastructure planning.

RQ3. How do AI method preferences vary across different social science disciplines?

AI adoption is unlikely to occur uniformly across research fields. Mapping disciplinary variation helps identify methodological specialisations, capability gaps and opportunities for interdisciplinary collaboration, while informing discipline-specific funding strategies, training programmes and responsible innovation initiatives.

RQ4. How has the distribution of AI method categories evolved over time?

Examining temporal trends reveals how AI capabilities diffuse through the research and innovation system, identifies emerging methodological trajectories and new research avenues, and enables UKRI to anticipate future capability requirements, governance challenges and strategic investment priorities.

RQ5. To what extent are responsible innovation principles evidenced within the design, implementation and reporting of AI-enabled social science projects?

While understanding AI adoption is essential, it is equally important to understand whether AI is being used responsibly. This question examines how principles such as anticipation, reflexivity, inclusivity, responsiveness, transparency and equitability are reflected within AI-enabled research projects. It provides evidence on whether responsible innovation has become embedded within everyday research practice and informs future governance, funding and policy development for AI-enabled research.

Collectively, these five questions move beyond documenting AI adoption towards understanding how AI is transforming the research and innovation system. Together, they provide evidence on where new methodological opportunities are emerging, how research practices are evolving, how responsible innovation is being operationalised, and how AI-enabled research can be governed in ways that support innovation, transparency, accountability, inclusivity and long-term research resilience.

Methodologically, the report develops a scalable computational framework for analysing AI adoption using funding-system data. AI-related projects funded since 2006 are identified by combining the GO-Science Emerging Technologies taxonomy with CWTS disciplinary classifications. A three-stage large language model (LLM)-assisted extraction pipeline identifies and classifies detailed AI methods together with responsible AI topics from project abstracts and associated project documentation. This enables the construction of a harmonised taxonomy encompassing machine learning, deep learning, foundation and generative AI, natural language processing, computer vision, reinforcement learning, simulation, statistical and Bayesian modelling, symbolic AI, multi-agent systems and trustworthy or privacy-enhancing AI. The resulting dataset enables systematic analysis of temporal diffusion patterns, disciplinary variation, methodological transparency, research outcome profiles and responsible innovation themes across the UKRI-funded social science portfolio.

The report makes three principal contributions. First, it develops and evaluates a scalable empirical framework for analysing AI adoption and responsible innovation practices using funding-system data and computational text analysis. Second, it advances conceptual understanding by examining how responsible innovation principles are operationalised through the adoption, governance and reporting of AI within publicly funded social science research. Third, it provides a robust empirical evidence base for researchers, funders and policymakers seeking to understand how AI-enabled research can be designed, governed and evaluated in ways that support innovation, transparency, inclusivity and long-term research resilience.

The findings presented in this report establish an empirical baseline for understanding AI adoption within UKRI-funded social science research. Future phases of the project will extend this work by examining how AI adoption relates to research performance, innovation outcomes, collaboration networks, funding trajectories and broader societal impacts. Subsequent analyses will apply a six-principle responsible innovation framework to evaluate whether AI-enabled projects demonstrate anticipation, reflexivity, inclusivity, responsiveness, transparency and equitability. This will move the analysis beyond identifying AI methods towards assessing how responsible innovation is evidenced and operationalised within publicly funded AI-enabled social science research, providing a richer understanding of how AI can contribute to research excellence while remaining socially responsible and trustworthy.

2. Data

2.1 Data Source

Our research uses the Gateway to Research Plus (GtR+) dataset, an enhanced, research-ready version of the UK Research and Innovation (UKRI) Gateway to Research (GtR) data. The original GtR data provide public information on UKRI-funded research and innovation projects from approximately 2006 onwards, covering grants, organisations, investigators, publications, and project outcomes. However, the raw GtR data has historically been difficult

to use for large-scale systematic analysis because it contains inconsistencies, missing contextual information, fragmented classifications, and limited interoperability across records. GtR+ data addresses these limitations by cleaning, enriching, standardising, and linking the original GtR data with additional datasets, metadata, and derived classifications. The enhanced dataset supports longitudinal and cross-disciplinary analysis of UKRI-funded research activity over time, enabling more robust identification of thematic trends, institutional collaborations, funding patterns, and emerging research domains.

2.2 Identifying AI-related Social Science Projects

The GtR+ dataset assigns each funded project one or more subject and technology labels, drawn from two independently applied taxonomies, the CWTS Research Fields taxonomy (Van Eck and Waltman, 2024) and the GO-Science Emerging Technologies taxonomy (Government Office for Science, 2025). The CWTS taxonomy classifies projects according to their disciplinary research area, covering four top-level fields within the GtR+ dataset: Physical Sciences, Life Sciences, Social Sciences, and Health Sciences. These are subdivided into 26 subfields in total. The GO-Science taxonomy classifies projects according to the emerging technology areas they engage with, spanning 17 top-level technology categories such as Energy, Geoengineering, and Artificial Intelligence, subdivided into 84 subfields in total. As shown in Table 1, each label assignment is described by the taxonomy source, a first-level label, a second-level subfield, and a rank giving the strength of that label relative to any others the project carries within the same taxonomy. A rank of 1 marks the label most strongly associated with the project.

Table 1. Structure of subject and technology label assignments in GtR+ dataset.

Variable	Description
reference	Unique project identifier assigned by the original funding body; never changes.
taxonomy	The classification system applied: either CWTS Research Fields or GO-Science Emerging Technologies.
taxonomy_label_l1	Top-level label from the relevant taxonomy, such as 'Life Sciences', 'Energy', etc.
taxonomy_label_l2	Subfield for the label listed in taxonomy_label_l1. For example, if taxonomy_label_l1 is 'Physics' then taxonomy_label_l2 is a subfield such as 'Astronomy'. Not all labels have a corresponding subfield.
rank	Relative rank for the different labels, where 1 is the strongest label. Some projects have multiple labels if they show strong alignment with multiple research fields, concepts, or emerging technologies.

This study focuses on projects that sit at the intersection of the two taxonomies: those with a first-level CWTS field of Social Sciences and a first-level GO-Science category of Artificial

Intelligence. A preliminary check confirmed that no substantive AI-related technology terms appear under other GO-Science top-level categories, supporting the use of this single category as the operational boundary for AI-related projects. Across the 125,446 projects in the full subject and technology dataset, 27,656 carry a Social Sciences label and 12,360 carry an Artificial Intelligence label, overlapping in 3,405 projects that carry both. After excluding projects that could not be matched to complete participant records, the final study corpus comprises 3,361 projects representing UKRI-funded social science research involving AI since 2006.

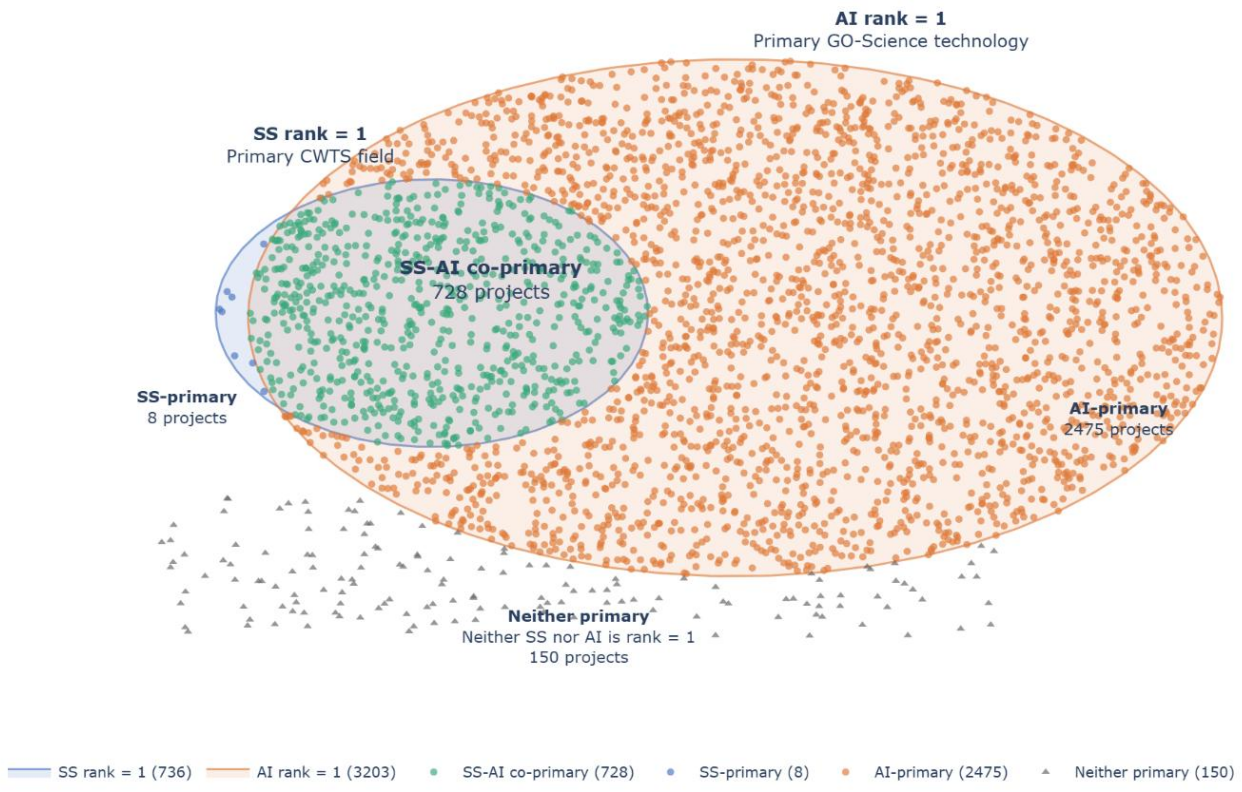

Figure 1: Funded social science projects involving AI.

Figure 1 illustrates the internal structure of this corpus by examining the rank-1 position of each label within the 3,361 projects, distinguishing projects in which social science is the primary CWTS research field (SS rank = 1) from those in which artificial intelligence is the primary GO-Science technology category (AI rank = 1). The SS-AI co-primary group (728 projects) lies at the intersection, comprising projects for which Social Sciences and Artificial Intelligence are both primary labels, and constitutes the analytically central cases of the corpus. The AI-primary group (2,475 projects), the largest of the four, carries Artificial Intelligence as its primary label while social science remains secondary, indicating projects where AI is central but the research sits primarily in areas such as Health Sciences, Engineering, or Computer Science. The SS-primary group (8 projects) carries Social Sciences as primary but AI only as secondary, while the Neither primary group (150 projects) holds both labels, neither ranked first. This distribution points to a focused but substantial overlap between social science and AI within the wider UKRI-funded research landscape. Social Sciences is the primary disciplinary label for only 736

projects (21.9%); for the remaining 78.1%, it appears as a secondary or co-occurring dimension, reflecting the interdisciplinary character of the corpus.

2.3 Trends and Duration of Social Science Projects

The project classification identifies three broad categories of UKRI-funded projects. The largest group consists of non-AI projects (111,972 projects), indicating that most funded research does not involve artificial intelligence technologies. AI-related projects are divided into two subgroups: AI interdisciplinary projects (11,362 projects), which combine AI with other disciplinary domains such as the social sciences, and pure AI projects (856 projects), which are primarily focused on AI itself. The substantially larger number of interdisciplinary AI projects suggests that AI is more commonly applied as an enabling methodology across research domains rather than pursued as a standalone research area.

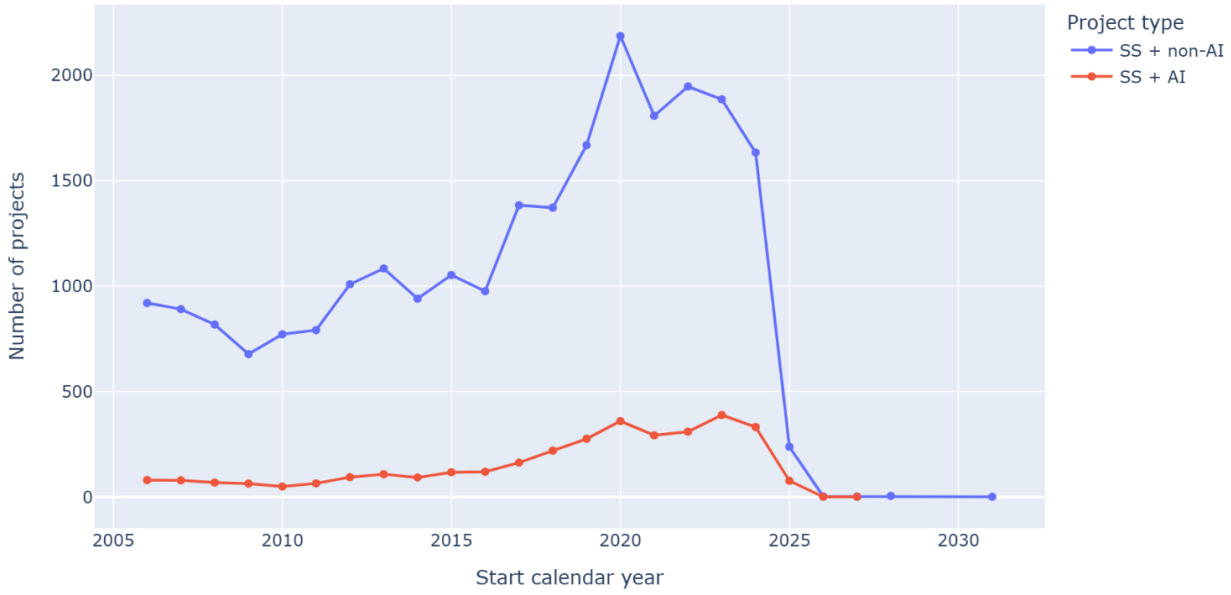

Figure 2: Annual distribution of AI-related and non-AI funded projects by start year.

Figure 2 compares the yearly number of AI-related and non-AI UKRI-funded projects based on project start year. Some UKRI projects are confirmed several years in advance; the GtR+ dataset used in this study was updated in December 2025. In addition, according to the GtR+ documentation, the underlying GtR system has an approximate six-month reporting lag. As a result, the recorded number of projects starting in 2025 and later years is likely incomplete and should not be interpreted as representing the full level of funding activity. AI-related projects show a clear upward trend over time, particularly from the mid-2010s onwards, reflecting the growing adoption of artificial intelligence across research domains. Non-AI projects remain substantially larger in number throughout the period. The sharp decline after 2024 is therefore most likely attributable to incomplete data coverage for recently funded or future projects within the current release of the dataset.

Figure 3 compares the duration distribution of AI-related and non-AI UKRI-funded projects. Most projects in both categories fall within the one-to-three-year and three-to-five-year duration ranges, indicating that medium-term funding periods dominate across the UKRI portfolio. AI-related projects show a similar duration profile to non-AI projects, although they appear slightly more concentrated within the three-to-five-year range. Very few projects in either category extend beyond ten years, suggesting that long-term funding commitments remain relatively uncommon within the dataset.

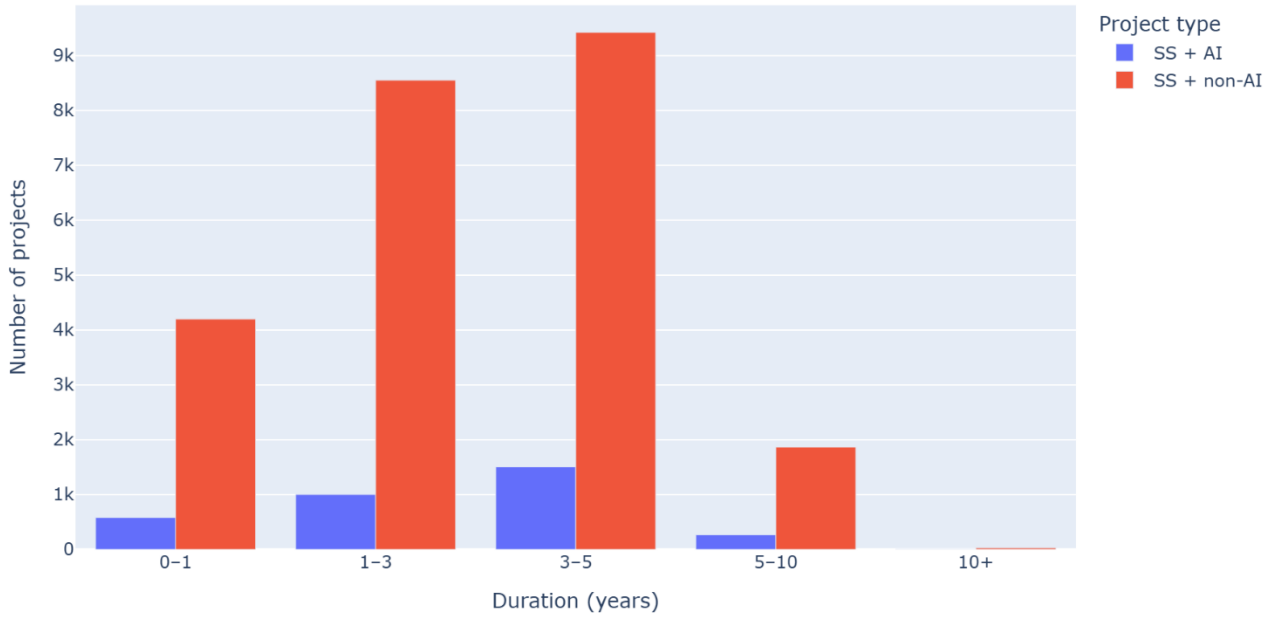

Figure 3: Project duration distribution for AI-related and non-AI projects.

2.4 Data Quality Review and Identified Issues

The data cleaning and quality review identified several issues that could affect the interpretation of the findings. These include zero or incomplete funding values, timing anomalies in start and end dates, repeated project-organisation records, inconsistent organisation identifiers, minor coding errors in organisation sector labels, and uneven coverage of recent projects and reported outcomes. These issues are not unusual in large administrative research-funding datasets, particularly where project, participant, organisation, funding, and outcome records are linked across multiple systems.

These issues have different implications for different parts of the analysis. Funding-related issues mainly affect project-level funding estimates; timing anomalies affect duration analysis and recent-year trend interpretation; organisation identifier inconsistencies affect institutional and sector-level summaries; and incomplete recent-year or outcome records affect conclusions about very recent projects and downstream research outputs. By contrast, the main AI method and responsible-AI topic analysis is less directly affected, because it is conducted at project level and relies primarily on project references and usable project text.

To address these issues, the report uses reference as the stable project identifier, conducts analysis at project level wherever possible, deduplicates project-organisation records before constructing funding measures, flags anomalous timing records, corrects unambiguous sector-label errors, and interprets 2025 and later years cautiously because of reporting lags and advance project confirmation practices. Projects without sufficient usable text are excluded from the detailed Stage 1 extraction, resulting in a cleaned corpus of 3,287 AI-related social science projects. Further details on the issues identified and the treatment applied are provided in Appendix A.

3. LLM-based AI Method Extraction and Classification

This report employs a three-stage analytical pipeline that combines automated extraction, structured human validation and offline taxonomy classification. This pipeline is designed to address the limitations of the raw GtR+ classifications, which provide only relatively broad AI sub-categories and do not capture the methodological detail required to examine patterns of AI adoption and diffusion across research domains. By integrating large language model (LLM)-based extraction with a fixed reference taxonomy and targeted human review, the approach enables scalable yet interpretable identification of AI methods and responsible AI topics from unstructured project text.

3.1 Stage 1: Automated Extraction

In the first stage, a prompted DeepSeek-V3 model processes each project's title, abstract, technical summary and potential impact statement, producing a structured record for each sentence. Rather than generating a single project-level output, the model performs sentence-level analysis, extracting terms relating to AI methods and responsible AI topics verbatim from the source text. For each term identified, the model assigns a stance indicating its relationship to the project's activities: whether the project applies the method or addresses the concern in its own conduct, examines it as an object of inquiry, or mentions it solely as contextual background. This approach maintains a clearly defined extraction task and promotes consistency, while reducing the likelihood that the model will overinterpret broad expressions or classify similar projects inconsistently. Model parameters were held constant across the corpus to ensure a consistent extraction process.

From these sentence-level records, a project-level classification of AI disclosure is derived. This distinguishes projects that explicitly name one or more specific AI methods from projects that reference AI only in broad or generic terms without specifying a concrete method, such as "AI-powered system" or "data-driven approach". Where a project claims AI use through such generic terms but names no specific method, it is additionally flagged as method-undisclosed. A project is credited with a method only where its own text presents that method as something the project itself uses. This approach is consistent with recent evidence showing that large language models can function as reliable automated coders for social science text annotation tasks when supported by detailed coding instructions (Gilardi, Alizadeh and Kubli, 2023). Full

details of the extraction configuration, including the model settings and worked examples, are provided in Appendix B.

3.2 Stage 2: Human Validation

In the second stage, human validation is used to check the consistency and accuracy of the automated extraction. The validation focuses on whether extracted terms are actually present in the project text, whether they are specific enough to count as AI methods, and whether the assigned stance correctly reflects the role of the term in the project. An initial pilot sample is reviewed to refine the extraction prompt and clarify decision rules. Particular attention is given to borderline cases, such as broad AI references without a named method, statistical or computational terms that may or may not be used as AI methods, and responsible AI terms that may appear either as project concerns or as background discussion.

After full-corpus extraction, records requiring adjustment are documented in a sentence-level repair log. This log records the term, source sentence, issue identified, and correction applied. The final validated output is then used for project-level disclosure coding and offline taxonomy classification.

3.3 Stage 3: Taxonomy Classification

In the third stage, the extracted and validated AI method terms are mapped to the harmonised taxonomy presented in Table 2 using an offline reference table. Within this table, each canonical method term is assigned either to one of ten top-level AI method categories or to one of four exclusion sets: umbrella terms, generic method phrases, measurement and data-collection instruments, and noise. As classification is conducted offline using a fixed reference table, rather than by the extraction model, all category assignments are fully auditable and may be revised without repeating the extraction process. Terms that cannot be matched to the reference table are logged and used to extend it iteratively.

In parallel, the extracted responsible AI concern terms are organised into concern families using a comparable offline procedure. Recurring clusters of terms are identified within the corpus, and their conceptual boundaries are refined with reference to established responsible innovation and AI ethics frameworks.

The definitions underpinning the AI method taxonomy draw on the ACM Computing Classification System (Association for Computing Machinery, 2012), the GO-Science Emerging Technologies taxonomy (Government Office for Science, 2025) and the taxonomy of AI proposed by Russell and Norvig (2022). These sources are supplemented by researcher judgement grounded in the computational social science literature (Evans and Aceves, 2016; Lazer et al., 2009; Lazer et al., 2020).

Table 2: AI method taxonomy and representative sub-families used for project classification.

Category	Representative sub-families
Machine Learning	Ensemble methods (random forest, XGBoost); linear models (SVM, logistic regression); clustering (k-means, DBSCAN); dimensionality reduction (PCA, UMAP); transfer learning; active learning; evolutionary computation (genetic algorithm, particle swarm optimisation)
Deep Learning & Neural Networks	Convolutional networks (CNN, ResNet, U-Net); recurrent networks (LSTM, GRU); transformers (BERT, ViT); graph neural networks (GNN, GCN); generative architectures (VAE, normalising flow); contrastive learning (CLIP, SimCLR)
Foundation Models & Generative AI	Autoregressive language models (GPT, LLaMA, Claude); diffusion models; vision-language models; prompting and alignment techniques (RAG, RLHF, LoRA, chain-of-thought)
Natural Language Processing	Topic modelling (LDA); word embeddings (word2vec, GloVe); sequence labelling (NER, POS tagging); text classification; sentiment and stance detection; machine translation; dialogue systems
Computer Vision & Signal Processing	Object detection; image segmentation; facial recognition; medical imaging; speech recognition (ASR); audio feature extraction; EEG signal processing
Reinforcement Learning	Value-based methods (Q-learning, DQN); policy gradient methods (PPO, actor-critic); multi-agent RL (MARL); inverse RL; offline RL; bandit algorithms
Symbolic AI / KRR	Knowledge representation and reasoning, knowledge graphs, ontologies, formal verification, logic programming, planning and constraint-based methods.
Multi-Agent & Synthetic Environments	Agent-based models (ABM); system dynamics; microsimulation; discrete event simulation; digital twins; cognitive models
Statistical & Other Computational Methods	Bayesian inference (MCMC, variational inference); Gaussian processes; causal inference; hidden Markov models; survival analysis; structural equation modelling; uncertainty quantification
Trustworthy & Privacy-Enhancing Computing	Post-hoc explanation (SHAP, LIME, Grad-CAM); federated learning; differential privacy; algorithmic fairness; adversarial robustness; human-in-the-loop methods

4. Key Findings

This section presents the main findings from the cleaned Stage 1 corpus of 3,287 social science projects involving AI.

Figure 4 presents the distribution of AI method categories identified from project abstracts. Machine Learning is the most prevalent category, appearing in 634 projects, substantially exceeding all other categories. Statistical / Other Computational Methods form the second-largest category, appearing in 228 projects, indicating the continued importance of statistical modelling, Bayesian/probabilistic reasoning, simulation-oriented analysis and wider computational methods within AI-related social science research.

4.1 Concentration and Diversity in AI Method Adoption

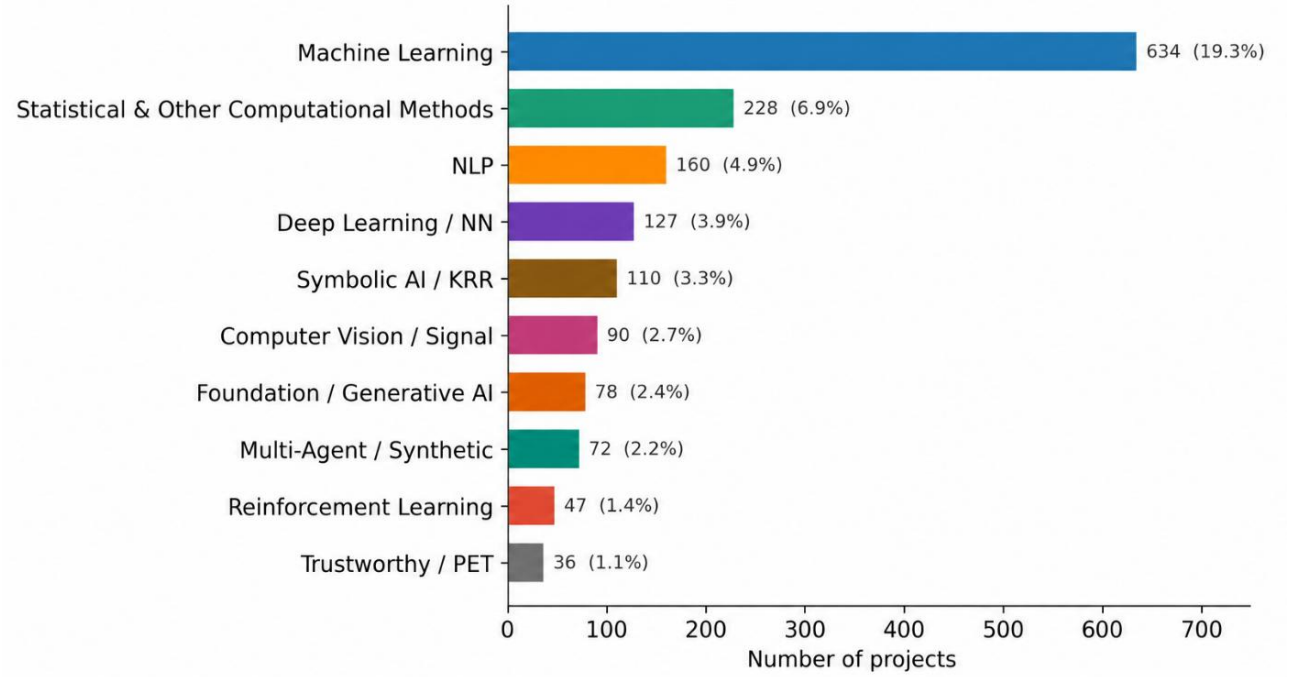

Figure 4: Distribution of AI method categories across AI-related social science projects.

Natural Language Processing appears in 160 projects, followed by Deep Learning / Neural Networks in 127 projects. These categories suggest a substantial but more specialised layer of language- and representation-based methods. Symbolic AI / Knowledge Representation and Reasoning is also visible, appearing in 110 projects, which indicates that rule-based, semantic and knowledge-structured approaches remain part of the methodological landscape.

Other categories appear less frequently but demonstrate the diversity of AI adoption across the corpus. Computer Vision / Signal Processing appears in 90 projects, Foundation / Generative AI in 78 projects, Multi-Agent / Synthetic Environments in 72 projects, Reinforcement Learning in 47 projects, and Trustworthy / Privacy-Enhancing Computing in 36 projects. The relatively smaller number of Foundation / Generative AI projects likely reflects the

recent emergence of these techniques compared with more established machine learning and statistical-computational approaches. Overall, the figure suggests that AI adoption within UKRI-funded projects is broad and methodologically heterogeneous, although strongly centred around machine learning and statistical modelling paradigms.

4.2 Evolution of Established and Emerging AI Methods

Figure 5 illustrates the temporal evolution of AI method categories identified from project abstracts from 2006 to 2024. The year 2025 is excluded from the trend analysis because the current GtR+ release has incomplete coverage for recently funded projects due to reporting lags and advance project confirmation practices within UKRI datasets. The figure suggests three broad methodological phases in the evolution of AI-related research within the UKRI-funded landscape.

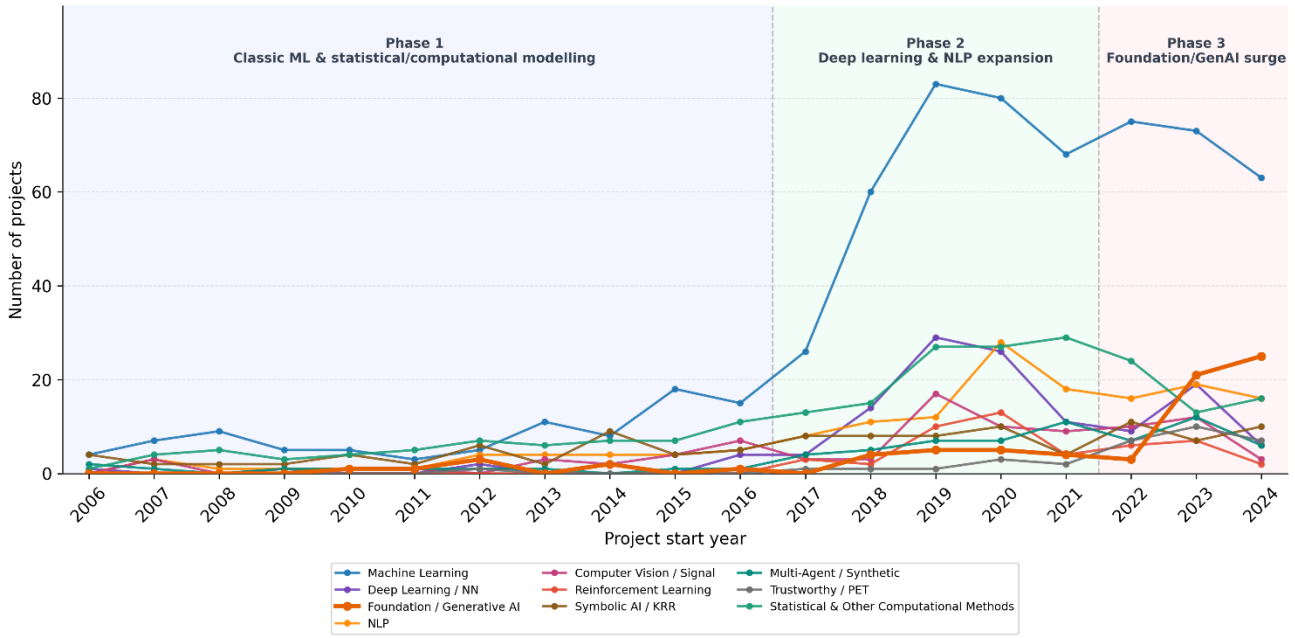

Figure 5: Temporal evolution of AI method categories in UKRI-funded projects.

Phase 1 (2006–2016) is characterised by relatively low-volume and gradual AI method adoption. Machine learning is already present during this period, but it does not yet dominate the funding landscape at the scale observed later. Statistical & Other Computational Methods show steady continuity, indicating that early AI-related social science projects were strongly connected to established quantitative, statistical, Bayesian/probabilistic and computational modelling traditions. Symbolic AI / KRR, NLP and computer vision/signal processing also appear intermittently, suggesting a methodologically diverse but still relatively dispersed early ecosystem.

Phase 2 (2017–2021) marks a clear expansion phase. Machine learning rises sharply and becomes the dominant methodological category in absolute project counts, reaching its highest

levels around 2019–2020. Deep learning and NLP also become more visible after 2018, reflecting the wider diffusion of neural architectures and large-scale text analysis into applied social science research. During this phase, AI adoption appears less experimental and more infrastructural: machine learning increasingly functions as a general-purpose analytical toolkit, while deep learning and NLP extend the capacity of projects to work with images, language, communication data and other high-dimensional materials. At the same time, statistical and computational modelling methods remain substantial, suggesting that newer AI approaches are layered onto existing computational social science infrastructures rather than replacing them.

Phase 3 (2022–2024) is characterised by the emergence of Foundation / Generative AI, particularly techniques associated with large language models and generative AI systems. Although this category remains smaller than machine learning in absolute terms, it shows the clearest recent upward movement, rising from 3 projects in 2022 to 25 projects in 2024. This pattern suggests that LLM-related and generative AI approaches are beginning to enter the UKRI-funded social science portfolio as a distinct methodological frontier. The rise is not simply another incremental increase within machine learning; rather, it points to a new layer of AI use centred on foundation models, generative systems, language interfaces and synthetic content generation. At the same time, earlier paradigms remain present: machine learning continues to be the largest method family, while NLP, symbolic AI/KRR, statistical-computational methods and other specialised categories continue to appear across the period.

Overall, Figure 5 suggests that new AI paradigms do not fully replace earlier methodological traditions. Instead, the UKRI-funded AI research ecosystem becomes increasingly layered: statistical, Bayesian/probabilistic and computational modelling methods remain stable; machine learning persists as the central methodological backbone; deep learning and NLP expand applied analytical capabilities; and Foundation / Generative AI, with LLMs as its most visible recent expression, introduces a new post-2022 generation of AI methods. The overall trajectory therefore reflects a transition from statistical and model-driven AI towards neural, representation-driven and increasingly generative foundation-model approaches, while retaining continuity with established quantitative social science methodologies.

Figure 6 presents a word cloud of emerging canonical AI methods and related concepts identified in project abstracts from 2020 onwards. The visualisation highlights the growing prominence of foundation-model and generative AI terminology within the UKRI-funded social science research ecosystem. Terms such as Large Language Model, Generative AI, ChatGPT, foundation model and foundation models appear prominently, indicating that LLM-based and generative approaches are becoming a visible methodological frontier in recent AI-enabled research.

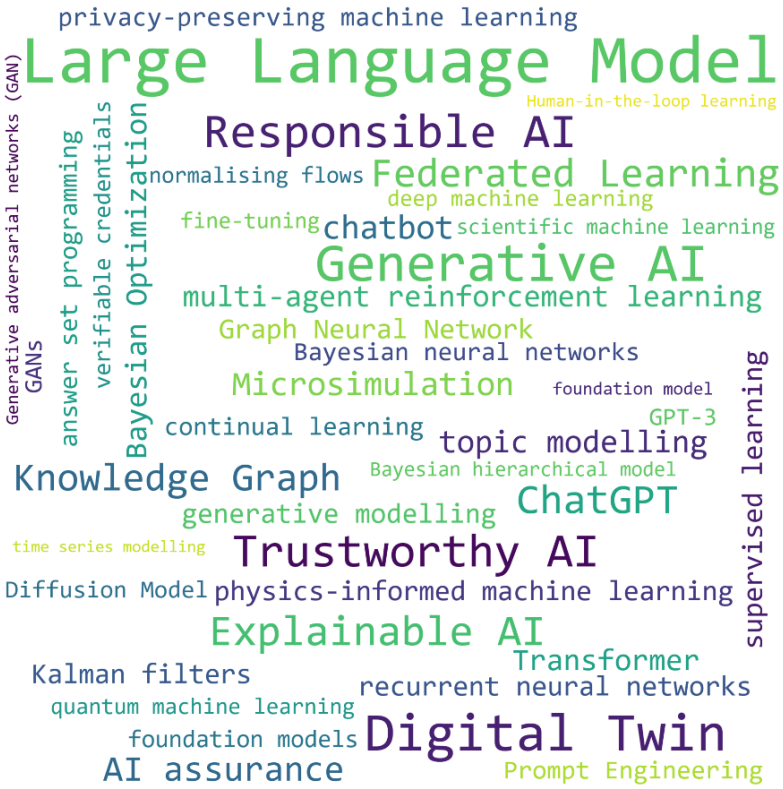

Figure 6: Emerging canonical AI methods in UKRI-funded social science projects from 2020 onwards.

It also shows the increasing salience of governance-, trust- and assurance-oriented concepts. Terms such as Responsible AI, Trustworthy AI, AI assurance, Federated Learning, Explainable AI, privacy-preserving machine learning and verifiable credentials suggest that recent AI method language is not only capability-oriented but is also increasingly connected to questions of privacy, transparency, accountability, security and responsible deployment.

At the same time, the word cloud points to a broadening set of specialised technical methods. Terms such as Transformer, Diffusion Model, Prompt Engineering, Graph Neural Network, multi-agent reinforcement learning, recurrent neural networks, Bayesian neural networks, GANs and normalising flows indicate the emergence of more technically differentiated AI paradigms. Knowledge Graph, answer set programming and other symbolic or structured-representation terms also remain visible, showing that recent methodological innovation is not limited to neural or generative approaches.

Several concepts bridge AI with established computational social science and modelling traditions. Digital Twin, Microsimulation, Bayesian optimisation, Bayesian hierarchical model, Kalman filters, time series modelling and physics-informed machine learning suggest that newer AI methods are being layered onto established modelling, simulation and statistical traditions rather than replacing them entirely. Overall, Figure 6 illustrates a post-2020 shift towards generative, foundation-model and governance-oriented AI, while also showing continuity with broader computational social science methods.

4.3 AI Method Preferences Across Social Science Fields

Figure 7 compares the distribution of AI method categories across major social science fields using row percentages. The results indicate substantial methodological variation between disciplines, suggesting that AI adoption within the social sciences is highly field-dependent rather than homogeneous. Machine Learning is the most widely used category across nearly all fields, particularly in Decision Sciences (26.8%), Business and Management (25.7%), and Economics and Finance (24.0%), reflecting the strong orientation of these disciplines towards predictive analytics and quantitative modelling.

Decision Sciences is strongly associated with Bayesian and Probabilistic methods (21.0%), reflecting its quantitative decision-analysis tradition. Economics and Finance exhibits the most technically intensive profile, with the highest shares of Deep Learning (12.5%) and Reinforcement Learning (9.4%), suggesting stronger engagement with optimisation, forecasting, and algorithmic decision systems. Arts and Humanities is distinguished by comparatively high levels of NLP (15.6%) and Computer Vision / Signal Processing methods (10.3%), indicating the growing integration of computational humanities and multimodal analytical approaches.

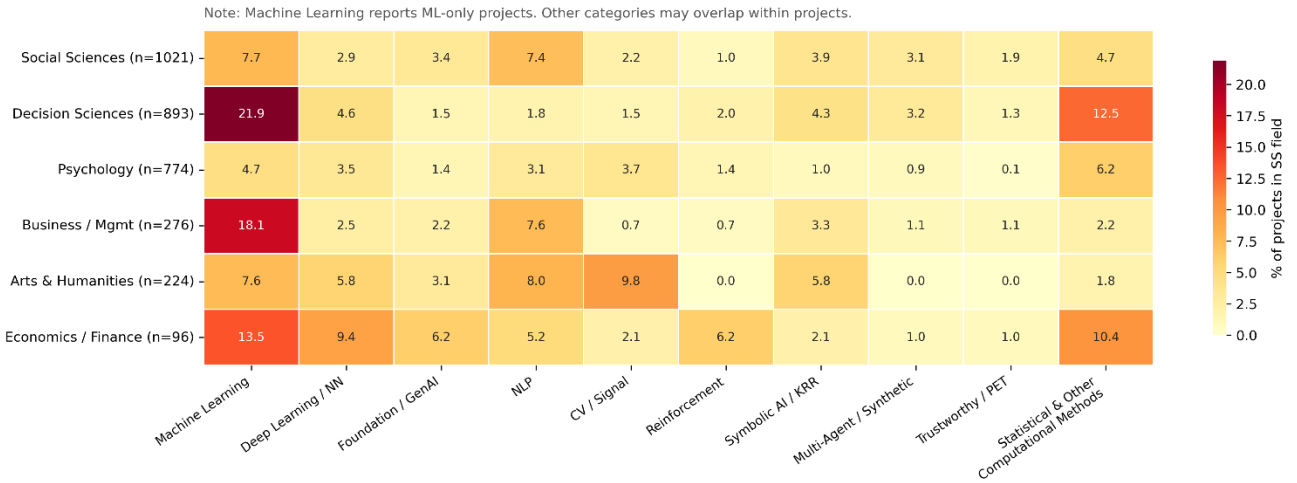

Figure 7: AI method preferences across social science fields.

Figure 7 shows how AI method categories vary across major social science fields, measured as the percentage of projects within each field. The “Social Sciences” category represents projects that could not be more precisely distinguished into the other detailed social science subfields based on the SS_Field variable labels. It is also the largest field in the dataset (n = 1,021) and displays a relatively dispersed methodological profile, combining Machine Learning-only projects (7.7%), NLP (7.4%), Statistical & Other Computational Methods (4.7%), Symbolic AI / KRR (3.9%), Foundation / GenAI (3.4%), and Multi-Agent / Synthetic approaches (3.1%).

The heatmap suggests that AI method adoption is strongly shaped by disciplinary context. Decision Sciences shows the strongest concentration of Machine Learning-only projects

(21.9%) and also has a high share of Statistical & Other Computational Methods (12.5%), consistent with its quantitative and optimisation-oriented traditions. Business / Management also has a strong Machine Learning profile (18.1%) and comparatively high NLP use (7.6%), suggesting a focus on prediction, decision support, organisational data and text-based business analytics. Arts & Humanities stands out for Computer Vision / Signal Processing (9.8%), NLP (8.0%), Deep Learning / NN (5.8%), and Symbolic AI / KRR (5.8%), reflecting the importance of images, language, archives, cultural data, and structured knowledge representation in this field.

Economics / Finance has a smaller sample (n = 96), but shows relatively high shares of Machine Learning-only projects (13.5%), Statistical & Other Computational Methods (10.4%), Deep Learning / NN (9.4%), Foundation / GenAI (6.2%), and Reinforcement Learning (6.2%). Psychology shows a more moderate profile, with Statistical & Other Computational Methods (6.2%), Machine Learning-only projects (4.7%), Computer Vision / Signal Processing (3.7%), Deep Learning / NN (3.5%), and NLP (3.1%). Trustworthy / PET methods remain comparatively rare across all fields, though they are most visible in the broad Social Sciences category (1.9%).

Overall, Figure 7 indicates that AI adoption does not follow a uniform cross-field pattern. Instead, different fields selectively adopt methods that align with their data structures, modelling traditions, and research questions: Decision Sciences and Economics lean towards machine learning and statistical-computational methods; Arts & Humanities shows stronger language, vision, and symbolic representation signals; and the broad Social Sciences category contains the most mixed and heterogeneous methodological profile.

4.4 Research Outcome Profiles by AI Method Categories

Figure 8 presents the average number of reported research outcomes per project across different AI method categories. The heatmap compares outcome profiles spanning publications, collaborations, engagement activities, software products, policy influence, intellectual property, and other forms of research impact. Darker shading indicates higher normalised intensity within each outcome category.

Several methodological patterns emerge from the figure. Computer Vision / Signal Processing has the strongest overall outcome profile, with the highest average number of publication rows (22.6), engagement activities (9.4), collaborations (3.0), further funding rows (2.0), research database/model outputs (1.6), and software/technical product rows (1.4). This suggests that CV / Signal projects are strongly associated with both conventional academic dissemination and more technical or applied outputs.

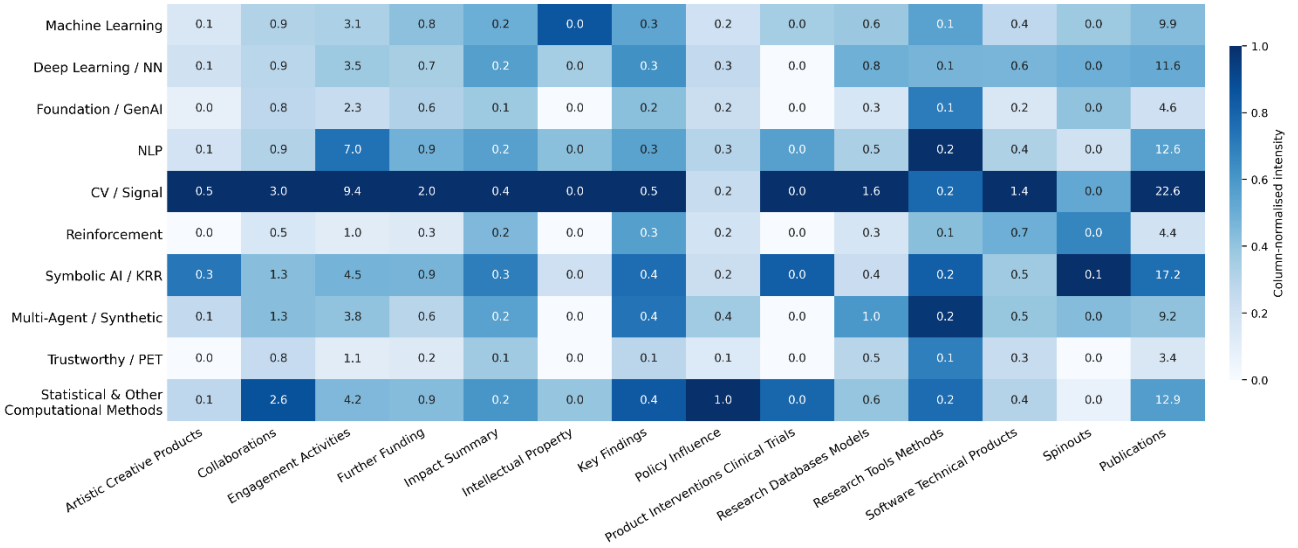

Figure 8: Average research outcomes per project across AI method categories.

Symbolic AI/KRR also shows a strong output profile, particularly in publications (17.2), engagement activities (4.5), research tools and methods, software and technical products, and key findings. This is consistent with the role of symbolic and knowledge-representation approaches in producing formal systems, structured resources, tools, and reusable technical infrastructure. Statistical & Other Computational Methods also display a broad outcome profile, with high publication intensity (12.9), collaborations (2.6), engagement activities (4.2), and the strongest policy influence score among the method categories (1.0), suggesting a close link between statistical-computational modelling and policy-relevant research outputs.

NLP and Deep Learning / NN projects show substantial publication and engagement activity. NLP records high publication output (12.6) and engagement activity (7.0), while Deep Learning/NN has high publication output (11.6) and moderate activity across collaborations, engagement, research databases/models, and software outputs. Machine Learning, as the largest and most general method family, shows a balanced but less concentrated profile, with publications (9.9), engagement activities (3.1), and moderate levels of collaboration, further funding, key findings, and research database/model outputs.

Foundation / GenAI and Trustworthy / PET currently show lower average outcome counts, with publication averages of 4.6 and 3.4 respectively. This likely reflects the more recent emergence of these categories within the funding portfolio: many projects may not yet have accumulated publications or formal outcomes at the same rate as older method families. Multi-Agent / Synthetic methods occupy an intermediate position, with moderate publications (9.2), engagement activities (3.8), and research database/model outputs (1.0). Reinforcement Learning shows a comparatively lower outcome profile overall, although it has some visibility in software/technical products.

Overall, Figure 8 suggests that outcome profiles differ substantially by AI method category. Publications remain the dominant reported output type across most categories, but the

strongest output intensity is not evenly distributed. Computer Vision / Signal Processing and Symbolic AI / KRR are especially output-intensive; Statistical & Other Computational Methods are comparatively policy- and collaboration-oriented; NLP and Deep Learning show strong academic and engagement profiles; and newer areas such as Foundation / GenAI and Trustworthy / PET appear less outcome-mature, consistent with their more recent growth in the corpus.

4.5 Responsible AI Topics and Their Coupling with AI Methods

The preceding findings describe how AI methods are distributed across the corpus, how they evolve over time, vary across social science fields, and relate to reported research outcomes. A further question is whether these projects also articulate responsible AI concerns, as the use of AI alone does not indicate that risks, fairness, transparency, stakeholder engagement, or governance have been considered.

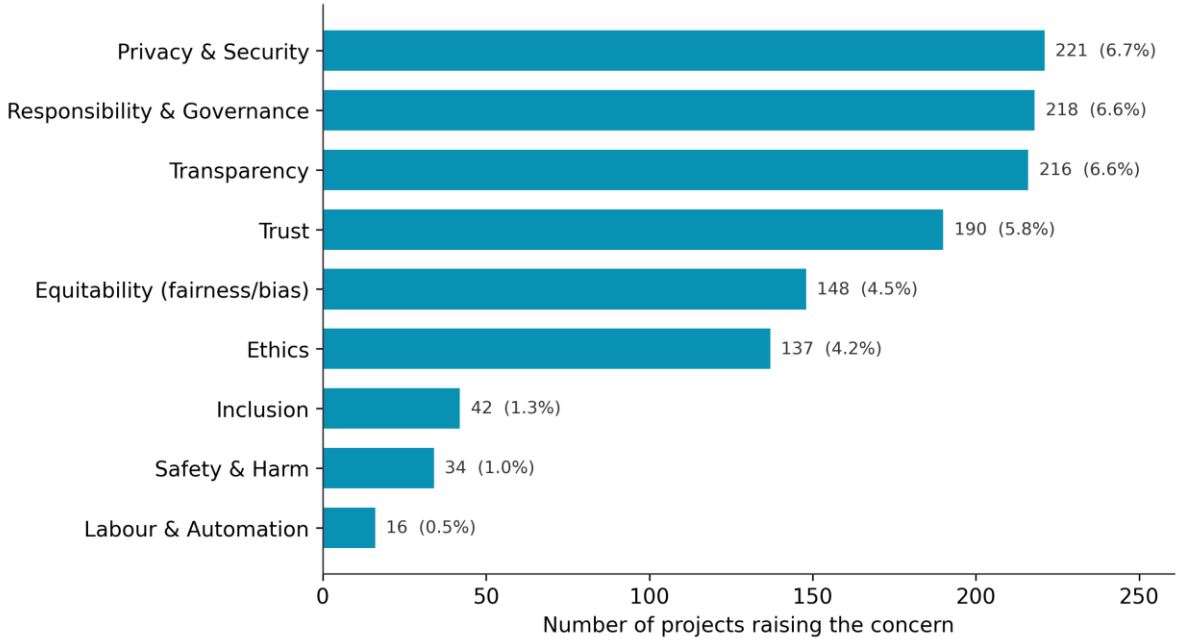

Figure 9: Responsible AI concern families in project texts.

Figure 9 summarises the responsible AI concern families identified in project texts. Responsible AI concern language is present in the corpus, but it is less common than method language. The largest concern families are Privacy & Security, Responsibility & Governance, Transparency, Trust, Equitability, and Ethics. These categories show that project texts contain a governance and legitimacy vocabulary as well as a technical vocabulary. However, the distribution is uneven: concerns related to inclusion, safety and harm, and labour or automation appear much less frequently, suggesting that responsible AI language is concentrated around a narrower set of recognisable governance and technical risk themes.

The relationship between AI method use and responsible AI concerns is uneven. Among the 1,053 projects that name a specific AI method, only 255 also articulate at least one responsible AI concern. Conversely, 368 projects articulate a responsible AI concern without naming a specific AI method. This means that responsible AI language is not simply the ethical counterpart of AI method adoption. Some projects use AI without articulating responsibility concerns in the project text, while others study or discuss responsible AI issues without naming a concrete AI method as part of their own research practice.

Figure 10: Responsible AI concern coupling by AI method category.

Figure 10 links the two dimensions directly by showing, for each AI method category, the share of projects that also mention each responsible AI concern family. The figure is descriptive rather than causal: it reports co-occurring textual signals, not evidence that a method causes a concern or that a project implements responsible innovation in practice. Nevertheless, the pattern is analytically useful because it shows where responsible AI language is most tightly coupled to particular technical methods.

The strongest coupling appears in Trustworthy & Privacy-Enhancing Computing. Projects in this category are much more likely than other method families to mention Privacy & Security, Trust, Responsibility & Governance, Transparency, and Equitability. This is the clearest evidence in the corpus that responsible AI concerns sometimes take a concrete methodological form: where responsibility becomes method, it often appears through privacy-preserving, trustworthy, explainable, assurance-oriented or governance-oriented technical approaches.

Other method families show more specific concern profiles. Deep Learning / Neural Networks projects are comparatively associated with transparency concerns, which is consistent with the explainability demands generated by opaque neural models. Symbolic AI / KRR projects show a relatively strong connection with trust and transparency, reflecting their association with

interpretable, rule-based or knowledge-structured reasoning. NLP projects show a modest association with equitability, which is consistent with wider concerns about bias, representation and fairness in language technologies. Foundation / GenAI projects also show an emerging responsibility profile, with transparency, trust and responsibility & governance visible alongside the rapid post-2022 rise of LLM-related terminology.

The absences are equally informative. Computer Vision / Signal Processing projects show comparatively weak coupling with responsible AI concern families, despite computer vision being strongly associated in the wider AI ethics literature with surveillance, recognition, classification and bias risks. This does not prove that such projects neglect responsibility; it may reflect proposal-writing conventions or the fact that responsibility mechanisms are not described in abstracts. However, it does identify a candidate blind spot for the next stage of analysis.

Overall, this section shows that responsible AI concerns are present but unevenly distributed. It is method-conditioned, concentrated in some technical families, and incomplete across the portfolio. This provides the empirical bridge to the next stage of the study: moving from topic detection to a six-principle responsible innovation framework that can distinguish rhetorical mention, research topic, and project practice.

5. Discussion

5.1 From AI Adoption to AI Capability

The findings show that AI adoption in UKRI-funded social science projects should not be understood simply as the uptake of a single new technology. It is better understood as the growth of a layered research capability. Machine learning forms the largest visible method family, but it sits alongside statistical modelling, Bayesian/probabilistic reasoning, simulation, symbolic AI, NLP, computer vision, multi-agent methods, trustworthy computing and, more recently, Foundation / Generative AI. This suggests that AI-enabled social science is not developing through replacement, where each new paradigm displaces the previous one. Instead, new methods are being added to an existing methodological base.

This layered pattern is important because it changes how AI capability should be interpreted. The rise of Foundation / Generative AI is one of the clearest recent shifts in the portfolio, but it is not the whole story. Much of the AI-related social science landscape continues to depend on established methods for modelling, classification, prediction, simulation, language analysis, decision support and uncertainty management. In this sense, AI capability is not concentrated only in frontier models. It is distributed across a broader set of methods, skills, data practices and disciplinary traditions.

The findings also suggest that social science is not only a user of AI methods. It is part of the environment in which AI becomes meaningful, usable and governable. Social science projects

help define the problems to which AI is applied, provide evidence about people and institutions, develop ways of evaluating social outcomes, and examine how technologies interact with public services, organisations, markets and communities. AI capability in this portfolio is therefore both technical and social: it depends on methods, but also on context, interpretation, governance and use.

5.2 AI-enabled Innovation and Responsible Practice

The findings have several implications for how the AI-related social science portfolio can be understood. First, the portfolio appears to be methodologically broad rather than concentrated around a single AI frontier. Foundation / Generative AI is emerging quickly, but established machine learning, statistical-computational methods, simulation, NLP and symbolic approaches remain important. This suggests that the strength of the portfolio lies partly in its diversity. Different methods support different kinds of research tasks, and newer methods often build on rather than replace older traditions.

Second, AI capability appears to be field-sensitive. Different social science fields use different data sources, methods and output pathways. Decision-oriented fields show stronger links with machine learning and statistical modelling; Arts and Humanities show stronger signals around language, vision and symbolic representation; Psychology links AI methods to behavioural and perceptual data; and the broad Social Sciences category combines several method families. This means that AI adoption is shaped by disciplinary needs rather than by a single general model of AI use.

Third, the relationship between AI and innovation is broader than the production of new technical tools. AI-enabled innovation can appear through new datasets, analytical methods, software, decision-support tools, policy evidence, public engagement, organisational learning and governance practices. Some of these outputs are visible in GtR+ outcome records, while others may be less directly captured. For example, a project that improves how sensitive data is analysed, how a public service evaluates AI, or how stakeholders are involved in AI design may contribute to innovation even if its outputs are not primarily commercial or technical.

Fourth, the responsible AI findings show that responsibility is present in the portfolio, but unevenly expressed. Responsible AI topics are visible around privacy, security, governance, transparency, trust, fairness and ethics. However, these topics are not consistently coupled with AI method use. Some projects name AI methods without articulating responsible AI concerns, while others discuss responsible AI issues without naming a specific method used in their own work. This suggests a distinction between AI as a method, AI as a topic of social concern, and responsible innovation as a project practice.

This distinction is important for interpreting the evidence. The presence of responsible AI vocabulary does not necessarily show that responsible innovation principles are embedded in project design. Equally, the absence of such vocabulary does not prove that a project lacks

responsible practice. Project abstracts are short, proposal-stage texts and may not describe every governance mechanism. The value of the current analysis is therefore to identify where responsibility is visible, where it is weakly articulated, and where further examination is needed.

5.3 Current Limitations

Several limitations should be kept in view. First, the project text is proposal-stage evidence. It records intentions, framing and projected contributions rather than completed practice. This means that the analysis can show how AI methods and responsible AI topics are described, but not whether all proposed activities were implemented in practice.

Second, GtR+ outcome records are uneven and affected by reporting practices. Outcome profiles should therefore be interpreted as recorded outputs rather than complete measures of project quality, productivity or impact. Newer method categories such as Foundation / Generative AI are also likely to be under-observed in long-term outcome measures because many associated projects are recent and have had less time to report outputs.

Third, the LLM-based extraction identifies candidate methods and responsible AI topics from project text, but it does not by itself determine the quality or depth of responsible innovation. A project may mention fairness, transparency or trust in a general way, study these issues as research topics, or build them into its own design. These are different forms of evidence and should not be treated as equivalent.

Finally, the current analysis is best understood as a Stage 1 portfolio map. It identifies the main AI method families, the rise of Foundation / Generative AI, the disciplinary variation in method use, the outcome profiles associated with different methods, and the uneven visibility of responsible AI topics. The next stage will need to examine responsible innovation more directly by assessing how projects articulate anticipation, reflexivity, inclusivity, responsiveness, transparency and equitability, and by distinguishing rhetorical mention from project-level practice.

6. A Proposed UKRI Framework for Evaluating Responsible AI in AI-enabled Research

The preceding analyses demonstrate that AI is becoming an increasingly important capability within UKRI-funded social science research. They also show that the relationship between AI adoption and responsible innovation is neither automatic nor consistent. While many projects describe sophisticated AI methods, relatively few explicitly articulate how issues such as transparency, fairness, governance, stakeholder engagement or societal impact are addressed. Conversely, some projects discuss responsible AI concerns without clearly describing the AI methods they employ. These findings suggest that methodological innovation and responsible innovation currently evolve along partially independent trajectories.

This evidence indicates that future evaluation should move beyond identifying whether AI is used and towards understanding how AI is designed, implemented, governed and evaluated throughout the research lifecycle. To support this transition, this report proposes a UKRI Framework for Evaluating Responsible AI in AI-enabled Research. The framework builds directly on the empirical findings of this study and provides a practical approach for embedding responsible innovation into proposal assessment, project delivery, impact evaluation and portfolio management. Rather than replacing existing UKRI assessment criteria relating to scientific excellence and impact, it provides an additional governance dimension for evaluating whether AI-enabled research is transparent, trustworthy, inclusive and socially responsible.

6.1 Conceptual Foundation

The proposed framework is grounded in the responsible innovation literature. Stilgoe, Owen and Macnaghten (2013) identify anticipation, reflexivity, inclusivity and responsiveness as the core dimensions of responsible innovation. Carbajal-Pina and Acur (2025) extend this framework by incorporating transparency and equitability, producing a six-principle structure that is particularly well suited to AI-enabled research. Within this report, responsible innovation provides the overarching governance framework, while responsible AI represents the practical operationalisation of these principles throughout the AI research lifecycle. Each principle is translated into observable evidence within project documentation. Anticipation relates to consideration of future risks and unintended consequences; reflexivity concerns assumptions, data quality and methodological limitations; inclusivity focuses on stakeholder participation; responsiveness considers mechanisms for adaptation and learning; transparency addresses documentation, explainability and traceability; and equitability evaluates fairness, accessibility and the distribution of benefits and risks.

6.2 Evidence and Coding Framework

The framework adopts a mixed deductive–inductive coding strategy. The six responsible innovation principles provide the overarching deductive analytical structure, while inductive refinement identifies recurring evidence patterns and clarifies their interpretation within project documentation. Rather than relying solely on keyword detection, the framework evaluates whether projects provide concrete evidence that responsible innovation principles are embedded in research design, implementation and governance.

For each principle, the coding distinguishes between three forms of evidence:

- Project practice, where the principle is operationalised within the project's own activities;
- Research object, where responsible AI is the subject of investigation; and
- General rhetoric, where the principle is referenced only at a broad or aspirational level.

This distinction enables the framework to move beyond simple mention counts towards assessing the extent to which responsible innovation is operationalised within funded research. Table 3 summarises the proposed coding framework.

Table 3: Six-principle responsible innovation framework for next-stage project coding.

Responsible innovation principle	AI-related focus	Evidence to look for
Anticipation	Possible risks, harms, misuse, uncertainty and longer-term consequences	Risk assessment, lifecycle risk, unintended consequences, future scenarios
Reflexivity	Awareness of assumptions, framing choices, data limits and model limitations	Discussion of data bias, model assumptions, uncertainty, proxy measures
Inclusivity	Involvement of stakeholders, users or affected groups	Co-design, public engagement, stakeholder workshops, community participation
Responsiveness	Ability to adapt, revise or respond to feedback and emerging issues	Monitoring, feedback loops, redress, revision of tools or methods
Transparency	Openness, documentation, traceability and explainability	Model documentation, explainable AI, open methods, audit trails
Equitability	Fairness, accessibility and distribution of benefits and risks	Fairness evaluation, subgroup analysis, accessibility, inclusion of underserved groups

6.3 Analytical Pipeline

The proposed evaluation pipeline builds directly upon the LLM-assisted extraction methodology developed in this report. Rather than identifying AI methods or responsible AI topics alone, the next stage evaluates how the six responsible innovation principles are articulated throughout project documentation.

Evidence is first identified at the sentence level and then aggregated into project-level profiles. Each piece of evidence is linked to the relevant responsible innovation principle, the strength of its articulation, and whether it represents project practice, research object or general rhetoric. These project profiles can subsequently be compared across AI method categories,

disciplinary fields, funding periods and research outcomes, enabling systematic assessment of responsible innovation across the UKRI portfolio.

6.4 Evaluation Across the AI Research Lifecycle

The framework adopts a whole-of-lifecycle perspective in which responsible innovation is evaluated continuously rather than at a single point during proposal assessment.

During **proposal development**, evaluation considers whether applicants justify the use of AI, anticipate societal implications, identify governance challenges and plan meaningful stakeholder engagement.

During **research design and implementation**, assessment focuses on data quality, methodological transparency, explainability, fairness, privacy protection, human oversight and mechanisms for monitoring emerging risks.

During **project delivery and impact**, evaluation examines stakeholder engagement, dissemination practices, methodological documentation, software and dataset transparency, policy engagement and demonstrated societal benefit.

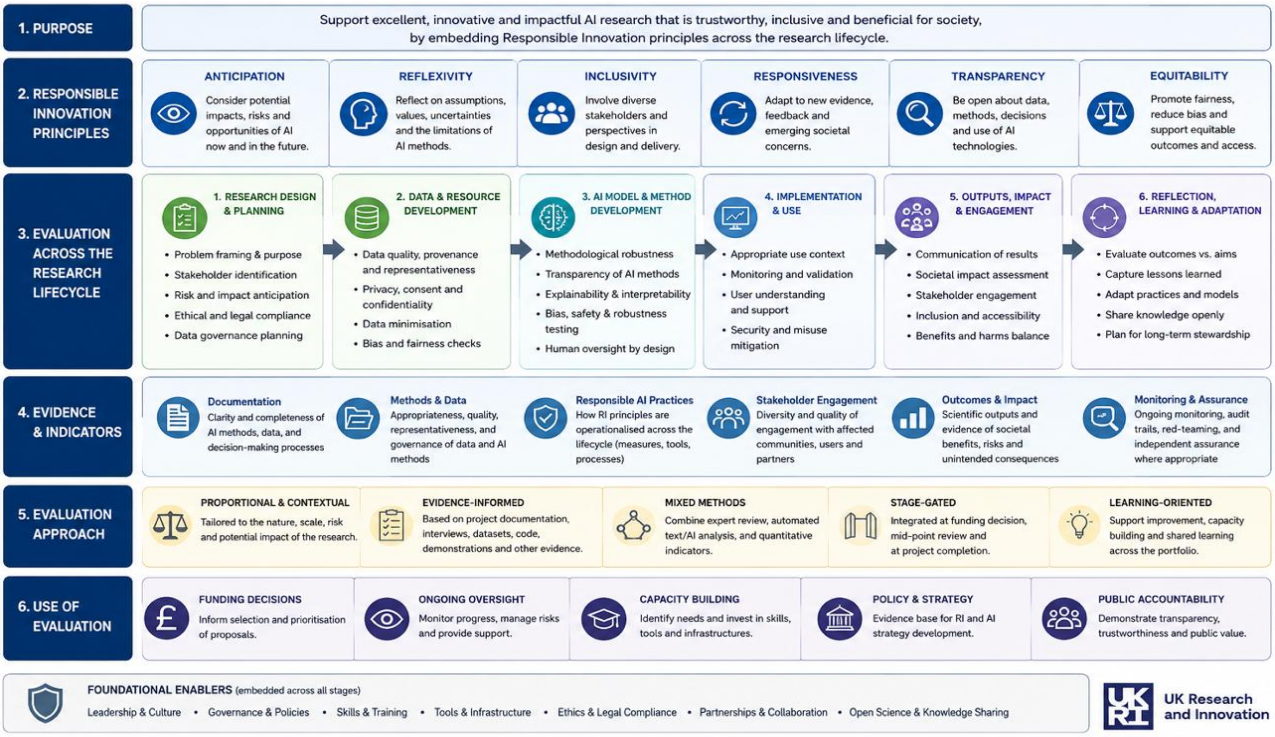

Figure 11: Proposed UKRI framework for evaluating Responsible AI in AI-enabled research.

Finally, **post-project learning** evaluates whether researchers reflect on lessons learned, adapt governance practices and contribute to organisational learning across the wider research and innovation system.

This lifecycle perspective embeds responsible innovation throughout the research process rather than treating it as a separate ethical compliance exercise. Figure 11 illustrates the proposed framework.

6.5 Evidence-based Responsible AI Evaluation

An important feature of the framework is its emphasis on evidence-based evaluation. Responsible innovation should be assessed using multiple forms of project evidence rather than relying exclusively on narrative statements within funding applications. These sources may include project documentation, methodological descriptions, datasets, software, stakeholder engagement activities, publications, impact reports and other research outputs generated throughout the project lifecycle.

Drawing on multiple evidence sources enables UKRI to evaluate how responsible innovation is implemented in practice, while reducing reliance on aspirational commitments made during proposal development.

6.6 Proportionate Governance

The framework also adopts a proportionate approach to governance. Different AI applications present different levels of opportunity and risk, and evaluation requirements should therefore reflect the nature of the research being undertaken. Projects involving sensitive personal data, foundation models or high-impact decision-making may require more comprehensive governance arrangements than lower-risk analytical applications.

A proportionate approach enables responsible innovation to strengthen research quality without imposing unnecessary administrative burdens on researchers.

6.7 Portfolio Monitoring and Strategic Decision-making

Beyond individual project evaluation, the framework provides UKRI with a strategic tool for monitoring the evolution of AI capability and responsible innovation across its research portfolio. Applied consistently, it would enable UKRI to monitor emerging AI methods, identify capability gaps, assess the maturity of responsible innovation practices, anticipate governance challenges and inform future funding priorities.

Longitudinal application of the framework would support evidence-based policy development, researcher capability building and infrastructure investment while strengthening transparency, accountability and public trust in AI-enabled research.

7. Conclusion and Strategic Implications

This report has examined the adoption of AI across UKRI-funded social science research, analysed the extent to which responsible innovation principles are reflected in AI-enabled research, and proposed an evidence-based framework for evaluating responsible AI throughout the research lifecycle. Together, these findings demonstrate that AI is reshaping both research practice and research governance, creating new opportunities while introducing new responsibilities for researchers, funders and policymakers. This concluding chapter summarises the principal findings, considers their implications for UKRI, and identifies priorities for future research and policy development.

7.1 Summary of Findings

This report has examined how AI is being positioned, described and implemented within UKRI-funded social science research. The evidence demonstrates that AI is no longer characterised by isolated applications or niche computational techniques, but has become an increasingly important methodological capability embedded across diverse areas of social science research. While machine learning remains the dominant AI paradigm, the portfolio also includes established statistical and probabilistic approaches, natural language processing, computer vision, symbolic AI, simulation, trustworthy and privacy-enhancing computing, and the rapid emergence of foundation and generative AI. Together, these developments demonstrate that AI is reshaping both the methods through which research is conducted and the questions that social scientists are increasingly able to investigate.

More broadly, the findings indicate that AI-enabled social science should be understood not simply as the adoption of new computational methods but as the emergence of a broader research capability. AI is enabling new forms of interdisciplinary collaboration, supporting novel approaches to evidence generation, accelerating knowledge discovery and creating opportunities to address increasingly complex societal challenges. At the same time, AI raises important questions concerning transparency, accountability, fairness, explainability, privacy, public trust and governance. The variation observed across disciplines further demonstrates that AI adoption is neither linear nor uniform but evolves through different methodological traditions, research cultures, data environments and pathways to societal impact.

7.2 Responsible AI and Strategic Implications

The analysis also provides important insights into the current state of responsible AI within UKRI-funded social science research. Although responsibility-related language is becoming increasingly visible within project documentation, its distribution remains uneven across AI methods and research domains. Stronger integration is evident in projects focused on trustworthy and privacy-enhancing computing, whereas many other AI method families demonstrate comparatively limited engagement with responsible innovation principles. Equally, a substantial number of projects discuss governance and ethics without clearly

describing the AI methods they employ. These findings suggest that methodological innovation and responsible innovation currently evolve along partially independent trajectories, indicating that greater integration of responsible research practices remains an important challenge for AI-enabled research.

These findings have important implications for UKRI. As AI becomes an increasingly integral component of publicly funded research, future funding, peer review and evaluation processes will need to assess not only whether AI methods are scientifically appropriate but also whether they are implemented transparently, responsibly and in ways that align with the principles of responsible innovation. The UKRI responsible innovation framework proposed in this report provides a practical, evidence-based approach to supporting this objective. By embedding anticipation, reflexivity, inclusivity, responsiveness, transparency and equitability throughout the research lifecycle, the framework moves beyond viewing responsible AI as a standalone ethical consideration and instead integrates responsible practice into proposal assessment, research design, project implementation, impact evaluation and organisational learning. Used alongside existing UKRI guidance, it offers a structured mechanism for strengthening research quality, improving consistency in evaluation and supporting trustworthy AI-enabled innovation across the research portfolio.

7.3 Future Priorities

The rapid evolution of foundation models, multimodal AI and increasingly autonomous research assistants will continue to reshape the social science research landscape. Future investment should therefore focus not only on expanding AI capability but also on strengthening researchers' capacity to deploy these technologies transparently, critically and responsibly. Greater attention is also needed to responsible innovation dimensions that currently receive comparatively limited attention, including inclusion, equity, labour transformation, environmental sustainability and longer-term societal impacts.

Future research should also move beyond identifying the presence of responsible AI language towards evaluating how responsible innovation is operationalised throughout research design, methodological decision-making, project governance and research outcomes. Longitudinal studies would be particularly valuable for examining how responsible practices evolve throughout project lifecycles and whether they contribute to research quality, collaboration, reproducibility and societal impact. Applying the proposed framework across future funding rounds would also generate an evidence base for understanding how AI capabilities and responsible innovation mature over time, enabling UKRI to identify emerging opportunities, anticipate governance challenges and refine funding priorities using empirical evidence rather than relying solely on normative guidance.

Ultimately, this report provides one of the first large-scale empirical assessments of how artificial intelligence is being adopted across UKRI-funded social science research while simultaneously examining how responsible innovation principles are reflected within AI-

enabled research practice. By combining technology classification, large-scale text analysis, LLM-assisted method extraction and a practical responsible innovation evaluation framework, it establishes an evidence base for understanding how AI capabilities, research practices and governance evolve together. AI should therefore be understood not simply as a collection of computational methods, but as a strategic research capability that is transforming the social science research and innovation system. The framework proposed in this report provides UKRI with a practical foundation for supporting AI-enabled research that is scientifically excellent, transparent, trustworthy and socially beneficial while informing future funding, governance and policy development.

REFERENCES

Association for Computing Machinery (2012) *The 2012 ACM Computing Classification System*. Available at: https://dl.acm.org/ccs.

Carbajal-Pina, C. and Acur, N. (2025) 'From principles to practice: Responsible implementation of emerging technologies through innovation dilemmas', *IEEE Transactions on Engineering Management*, 72, pp. 18-28. doi:10.1109/tem.2024.3479413.

Evans, J.A. and Aceves, P. (2016) 'Machine translation: Mining text for social theory', *Annual Review of Sociology*, 42(1), pp. 21-50. doi:10.1146/annurev-soc-081715-074206.

Gilardi, F., Alizadeh, M. and Kubli, M. (2023) 'ChatGPT outperforms crowd workers for text-annotation tasks', *Proceedings of the National Academy of Sciences*, 120(30). doi:10.1073/pnas.2305016120.

Government Office for Science (2025) *GO-Science Emerging Technology Taxonomy*. Available at: https://foresightprojects.blog.gov.uk/wp-content/uploads/sites/191/2025/09/GO-Science-Emerging-Technology-Taxonomy.pdf.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2009) 'Computational social science', *Science*, 323(5915), pp. 721–723. doi:10.1126/science.1167742.

Lazer, D.M.J., Pentland, A., Watts, D.J., Aral, S., Brewer, D., Christakis, N., Contractor, N., Fowler, J., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2020) 'Computational social science: Obstacles and opportunities', *Science*, 369(6507), pp. 1060–1062. doi:10.1126/science.aaz8170.

Owen, R., Macnaghten, P. and Stilgoe, J. (2012) 'Responsible Research and Innovation: From Science in Society to Science for Society, with Society', *Science and Public Policy*, 39(6), pp. 751–760. doi:10.1093/scipol/scs093.

Russell, S. and Norvig, P. (2022) *Artificial Intelligence: A Modern Approach, 4th US ed.* Available at: https://aima.cs.berkeley.edu.

Stilgoe, J., Owen, R. and Macnaghten, P. (2013) 'Developing a framework for Responsible Innovation', *Research Policy*, 42(9), pp. 1568–1580. doi:10.1016/j.respol.2013.05.008.

Van Eck, N.J. and Waltman, L. (2024) 'An open approach for classifying research publications', Leiden Madtrics. Available at: https://www.leidenmadtrics.nl/articles/an-open-approach-for-classifying-research-publications.

APPENDIX

Appendix A. Summary of Data Issues

This appendix summarises the main data issues identified during preparation of the GtR+ dataset and how they were handled in the report. Overall, these checks support the use of GtR+ for portfolio-level analysis, while also clarifying where caution is needed. The main findings on AI method adoption, responsible AI topics, disciplinary variation and temporal change are based on the cleaned 3,287-project corpus.

Table 4: Data quality issues and treatment.

Issue identified	Why it matters	Treatment in the report
Zero funding values of projects	Zero funding may reflect reporting conventions, incomplete records, or future or advance-confirmed projects, not necessarily truly unfunded work.	Zero funding values are retained for project identification but treated cautiously in funding-related interpretation.
Anomalous timing of projects	A small number of records contain future dates, invalid end dates, negative durations, or unusually long durations.	Timing anomalies are flagged. Duration and recent-year trend analysis are interpreted cautiously.
Reporting lags of data	GtR/GtR+ has reporting lags, and some projects are confirmed in advance.	The report does not treat the decline after 2024 as evidence of reduced funding activity. 2025 and later years are excluded or interpreted with caution in trend analysis.
Unaligned organisation names and IDs	The same organisation may appear in slightly different ways across records.	Organisation-level summaries use cleaned organisation information where possible.
Organisation-sector typo	A few records contain a spelling error in the sector label for public-sector organisations.	The typo is corrected before sector-level interpretation.

Appendix B. Extraction Pipeline and Worked Examples

This appendix describes the extraction pipeline referred to in Section 3.1, covering the model settings, the extraction and stance rules, the aggregation from sentence-level to project-level results, and worked examples drawn from the corpus.

B.1 Model Settings and Extraction Task

Extraction was carried out with the DeepSeek-V3 model, accessed through its 'deepseek-chat' interface. The model's sampling temperature was fixed at zero throughout to minimise sampling variability and support consistency and reproducibility across the corpus. Each project was processed in a single request, with its full text held in context, so that the interpretation of any one sentence could draw on the project as a whole rather than being made in isolation. The model was instructed to return structured output only, and every response was checked against a fixed format before it was accepted.

The model was given a narrow and clearly bounded task. For each project it received the title as context and the project's sentences as a numbered list, each tagged with the field it came from, whether the abstract, the technical summary, or the impact statement. Working sentence by sentence, the model identified two kinds of expression and recorded them exactly as they appeared in the text: the names of AI methods, and the names of responsible AI concerns such as transparency, fairness or privacy. For each expression it also recorded a stance describing how that expression related to the project's own work. The model was not asked to decide which family a method belonged to, or which responsible innovation principle a concern reflected. Those judgements were made afterwards, offline, using fixed reference tables, as described in Sections 3.3 and 4.4. Keeping categorisation outside the model has an important consequence for the analysis: the step that matters most for the results is carried out against an inspectable reference table rather than inside the model, so it can be examined and revised without repeating the extraction.

B.2 Stance Rules and Project-level Aggregation

The stance separates a project's own methods and commitments from those that merely appear in its text. A method was recorded as "used" where the project applies or develops it, "studied" where the project examines the method itself, and "background" where it belongs to others or to the field in general. Concerns were treated the same way, with "practice" in place of "used": a concern is coded "practice" where the project commits to it in its own conduct, "study" where it is the project's research subject, and "background" where it is only context. The distinction between "practice" and "study" is central to the responsible innovation interpretation, since a project that studies fairness is not, on that basis alone, embedding fairness in its own work.

Sentence-level records are then combined for the project. A project counts as naming a specific AI method where at least one method is coded "used" in its own work; where it refers to AI only

through umbrella terms and names no specific method, it is recorded as claiming AI but leaving its method undisclosed. Methods coded "used" are mapped to their families, and concerns are grouped into families and summarised by their strongest stance.

The extraction rules were calibrated on a pilot sample that was manually checked before the full corpus was processed.

B.3 Worked Examples

The two projects shown in Tables 5 and 6 illustrate how sentence-level codes combine into a project-level result.

The first is a doctoral project on multimodal language models. The extraction distinguishes the models the project builds on from the methods it will itself use, and records interpretability as a subject of study rather than a commitment.

Table 5: Sentence-level method and topic extraction results, Example 1.

Sentence	Methods (stance)	Concerns (stance)
"LLMs have demonstrated a remarkable ability to generate text from images, text, audio and video."	LLMs (background)	—
"The project will examine different approaches for multimodal LLM-based NLP."	multimodal LLM-based NLP (use)	—
"The PhD will delve into LLM architectures, prompting engineering and interpretability."	LLM architectures (use); prompting engineering (use); interpretability (study)	interpretability (study)

Project-level result: names specific methods in the Generative AI and NLP families, and identifies a transparency concern as a research object, not practice. The language models named in the opening sentence do not count towards the project's own methods, because they are recorded as background.

The second example is an innovative project developing an interpretability tool. As the project did not specify the exact methods used when referring to "artificial intelligence" throughout, this case was coded as an undisclosed method claim.

Project-level result: claims AI use but discloses no specific method, because the AI it refers to is background or a general capability rather than a named method it applies. It nonetheless registers a transparency concern as practice, reflecting its commitment to build an explainability tool. This pairing of an undisclosed method with a genuine responsible AI commitment is exactly what the disclosure and stance coding is designed to capture.

Table 6: Sentence-level method and topic extraction results, Example 2.

Sentence	Methods (stance)	Concerns (stance)
"AI can revolutionise computer-aided engineering."	AI (background)	—
"Like other AI, it is hampered by the 'black box' dilemma."	AI (background)	black box dilemma (background)
"The project builds the world's first AI-explainability tool for complex 3D designs."	AI-explainability tool (use)	AI-explainability (practice)

Every response was validated against the expected format, and any incomplete record was repaired and logged for review rather than discarded. These repairs affected a small fraction of the corpus and did not materially change the reported distributions.

www.ircaucus.ac.uk

Email info@ircaucus.ac.uk Twitter @IRCaucus

INNOVATION & RESEARCH CAUCUS

Delivered with ESRC and Innovate UK