The Royal Institution
Science Lives Here

AI policy v4

STATUS: Approved

Policy Effective Date: October 2024

Policy Owner:
Katherine Mathieson, Director

Date approved: 24 September 2024

Next Review Date: Sept 2025

1 Background, definition, purpose

1.1 This policy has drawn on guidance developed to address AI ethics and governance by the Alan Turing Institute[1] and a range of policies, including those recommended by CivicAI.[2] Marvin Minsky[3], cognitive scientist and AI pioneer, defined AI as, 'Artificial Intelligence is the science of making computers do things that require intelligence when done by humans.' References to AI in this policy refer to Generative AI – see definitions in Section 5.

1.2 AI continues to evolve and its application and integration changes at speed. We recognise that this policy is a starting point. The ongoing impact of integration and deployment will result in policy revision, as it becomes apparent.

1.3 The **purpose** of this policy is to establish principles for the Ri's day-to-day use of AI. It aims to support positive use where it assists us to better achieve our charitable objectives and is in line with the Ri's values, as well as acknowledging and mitigating risks, wherever possible.

1.4 We recognise that the Ri has very limited agency or control over the development and application of AI; every published document or webpage can be scraped and reused. We may also be impacted by working with outsourced or other external partner, funder or government agency AI use or integration.

1.5 **Scope:** This policy applies to all staff and those working on behalf of the Ri. Where relevant, it may also apply to third party suppliers, contractors, and stakeholders. The policy applies to all types of AI, including integrated tools e.g. Microsoft 365 co-pilot, Canva, Photoshop, Edge. It may include any data that we extract, input and/or manipulate / process however used to produce / process / input text, images, video, audio, code, publications, or any other media.

1.6 **Governance:** the policy will be scrutinised by the Audit and Risk Committee, who will recommend it for Trustee approval at least annually. Review will take place whenever technology and our use develops, knowledge grows and/or opportunities and risks arise, that require policy direction or clarification.

2 **Principles:** The Ri has adopted and adapted the Alan Turing Institute FAST principles to reflect the structure, values, and activities of the Ri:

[1] https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf
[2] https://civicai.uk/p/issue-1-organisational-policies?utm_source=substack&utm_medium=email
[3] https://web.media.mit.edu/~minsky/

Fairness: we will

- **Maintain and reflect the Ri's core values at all times,** including Equity, Diversity, Inclusion and Accessibility (EDIA), safeguarding those at risk and challenging anti-discriminatory practices
- **Process data legally and fairly** and maintain a principle of 'do no harm'
- **Encourage use** where it supports / augments EDIA, fairness, staff wellbeing, personal development and/or advances our charitable aims, for example: streamlining or automating routine tasks / reducing workloads
- **Fact check** and **research sources,** mitigating against bias, misinformation, human error, and/or unreliable data wherever possible
- 'Think AI' **– is it fair to share?** considering how information may be used outside of the Ri before we use it
- Have the right **consent** in place to share data/information

Accountability: we will

- Ensure a **human remains in the loop.** Humans create and AI assists
- Be **individually** responsible and **accountable** for our work
- Be aware that **all data / information put in, can be taken out** and used outside of our control
- **Not use AI to make decisions** for us
- Use AI in accordance with legislation, relevant regulation, and guidance particularly Data Protection / UK-GDPR
- **Protect rights and freedoms** and not use AI where it could undermine fundamental rights or privacy
- Use reasonable endeavours to understand when data may be being processed/stored **outside of UK legal jurisdiction** and assess relevant risks; acknowledging that controls may be weaker
- Remember: **Safeguarding of children** and adults at risk remains paramount
- Input data or information into AI systems **only with appropriate consent** in place

Sustainability: we will

- Use AI to support the Ri's **overarching charitable purpose**, objectives and **reputation**
- Not knowingly **risk loss of trust** internally or externally
- Be aware of increased fraud risks when using AI
- Take care not to **isolate / exclude** certain groups of people
- Assess **real-world, long-term impacts** of AI use in the workplace including environmental impact
- Take all reasonable steps to ensure systems we employ are **safe, accurate, reliable, secure, and robust**
- **Put People first,** supporting staff and using AI where it enables and/or augments day-to-day work
- Consider **welfare and wellbeing** as essential and assess impact of AI use on all those working on behalf of, or engaging with, the Ri
- **Encourage use** of AI where it can enable human social engagement, socialising, and connection

Transparency: we will

- **Be open and transparent** about our use of AI
- **Declare our use of AI and cite sources where we would otherwise claim author or ownership or cite research sources**
- **Not claim ownership** of any AI produced content
- Be open to **scrutiny** and ready to explain our use and safeguards
- **Report incidents, events,** or challenges without undue delay
- **Not use AI to track or monitor staff** without risk and impact assessment, clarity of purpose, prior approval, declaration, and adequate data protection
- **Be open to opportunities** that may improve, assist and support staff to augment and/or advance how we meet our outcomes
- **Support positive use** that helps us meet our charitable objectives effectively, safely, and inclusively

3 Policy detail

3.1 The Ri is committed to data protection and protecting people's fundamental rights and freedoms. This may also apply to data belonging to our or stakeholders, including members, patrons, supporters, and volunteers. All staff must adhere to relevant legislation and regulation, when using AI. The following must not be knowingly entered into any AI systems without consent:

- sensitive or personal data including images and/or names of staff
- information / images that could identify any person
- original content not owned by the Ri
- commercially sensitive data about the Ri or our partners

3.2 All content produced using AI must uphold Ri values and actively promote our commitment to Equity, Diversity, Inclusion and Accessibility (EDIA). All material must be reviewed ensuring that we take all reasonable steps to mitigate against the use of stereotypical, biased, offensive, and/or discriminatory content.

3.3 All AI generated information / data must be fact and source checked. AI can produce information that is inaccurate, misleading, outdated, or wrong. It can be poorly designed and/or use authoritative language that disguises invented data or misinformation. There is a risk of loss of trust and reputation with our funders, staff, stakeholders, and partners if poor quality, damaging, or incorrect information is produced.

3.4 AI generated information / data that is used should be appropriate to the UK, its diverse culture, legislation, and social values. Any data that is used must derive from relevant datasets e.g. Office of National Statistics, UK Government, relevant peer reviewed or authoritative research.

3.5 All content generated by AI, used either in part or full must be evaluated before being shared or published, declared, identified and cited. i.e. Microsoft Co-pilot for Edge/Bing cites its references[4].

3.6 AI will not make decisions on behalf of the Ri. We recognise that decision automation bias carries risk of influencing human interpretation and damaging critical thinking and decision making. AI generated predictions will not be relied upon unless they have been independently tested and verified.

3.7 Staff are responsible for checking and documenting if external reports that could influence key decision making have used AI, e.g. to benchmark, analyse data or reach

[4] Microsoft Co-pilot for Edge and Bing as deployed 2024

conclusions / recommendations. This could influence the Ri's position on tender specifications, procurement, provider / supplier fees and/or the reliability of actual or potential conclusions or reports overall.

3.8 Staff will report incidents, events, or challenges in line with usual business practice and relevant and current legislation, regulation, and Ri policies, for example, data protection, safeguarding and/or cyber fraud.

3.9 The Ri is committed to safeguarding the rights, wellbeing, and safety of its staff. We will be transparent about staff monitoring or data collation that employs AI. Access to confidential data will be strictly controlled with appropriate safeguards.

3.10 All staff must be aware of the potential for AI to be hacked, used for criminal activity and/or support fraud, e.g. voice replication, phishing emails. Information that is put into initially secure AI products e.g. within Microsoft 365 boundaries may be compromised if plug-ins or extensions are used.

3.11 Informal transcripts of online meetings can be used as a record of proceedings, providing the use is declared to all at the outset of the meeting. They must not be shared outside of Ri systems but may be shared, password protected, with recognised external partners. If a transcript is used a warning about potential inaccuracies must be issued or documented. AI generated transcripts will not be used as a final and/or evidential record of confidential or evidence-based meetings.

3.12 Formal governance meeting minutes are not intended to be a transcript and transcripts may not be used. They may be used for reference. All names and data must be checked and not relied upon as evidence.

3.13 Complaints and sensitive matters must not be responded to wholly using AI.

4. Examples: how we may use AI (this list is not exhaustive)	
Corporate and external publications: Content intended for external readers. Produce written or other content including blog posts, 'explainers,' Ri-owned website and social media content. Funding bids, impact, and evaluation reports.	- design, style, proof reading - suggesting content and presentation enhancements - model different approaches - templates, examples, and ideas - support practice development - in support of AI events / talks / demos, e.g., - structure or template external facing policy / documents - video content production including creating AI imagery, including that created from text - image / audio manipulation / enhancement
Desk research/report writing: Committee and Board reports and briefings, developing business case, policy development	- archive and online searching - due diligence information gathering / funding information - policy templates, research, and structure: pooling and summarising relevant information - summarising large reports / legislation / regulation - analysis of financial / numerical data, - support report / business case / decision making - pooling and analysing large datasets (e.g. national or external research) - identifying diverse sources to aid research - offering a range of viewpoints or positions for consideration - gap checking (e.g. identifying essential points required in a document) - ask AI models to act as an expert to review content

Human Resources	- processing routine data to support staff management (e.g. annual leave entitlements, flexible working) - monitoring (e.g. automating routine tasks)
Partnerships	- joint working with data, information sharing and analysis (e.g. ticketing) - automated outsourcing e.g., payroll, HR systems - developing / use of Salesforce
Personal data processing	- ticketing, membership records - bulk mail / communications, - volunteer / Trustee and Committee member checks - DBS (where used by external providers / Government) - payroll data - facial recognition may be used by police or security forces when reviewing CCTV in detection of crime
Policies/ guidance	- producing or reviewing internal policy / guidance - pooling / comparing relevant examples - templates / structure - summarising legislation / regulation or larger pieces of information
Recruitment	- screening applications - outsourced (agency) use - screening test results, where used
System procurement /development /upgrade	- includes third party and outsourcing (e.g. HR processing) - IT development and integration (e.g. Microsoft co-pilot) - Use of code (e.g. Salesforce)

5. Definitions

Algorithm: set of step-by-step instructions. In artificial intelligence, an algorithm tells the machine how to find answers to a question or solutions to a problem.[5]

Artificial intelligence: defined 1.1

Generative AI: broad label describing any type of artificial intelligence used to create new text, images, video, audio, or code. Large Language Models (LLMS) are part of this category and produce text outputs. It also covers AI that generates images based on text.[6]

Machine Learning: the process of computers improving their own ability to carry out tasks by analysing new data, without a human needing to give instructions in the form of a programme, or the study of creating and using computer systems that can do this.[7]

[5] https://www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard

[6] Summary of definition used in https://www.gov.uk/government/publications/guidance-to-civil-servants-on-use-of-generative-ai

[7] https://dictionary.cambridge.org/