DARE UK (Data and Analytics Research Environments UK)

DARE UK Interest Group (IG) Charter

Name of proposed Interest Group: Evaluation of Automated Output Checking and AI Model Risk Assessment

The WHY

Introduction:

The DARE Driver project Semi-Automated Checking of Research outputs (SACRO), and before it the DARE Sprint exemplar Guidelines and Resources for AI Model Access from Trusted Research Environments (GRAIMatter) delivered a suite of guidelines and tools for Output Statistical Disclosure Control (OSDC). These have established considerable interest and appetite within the UK and international community, that we now need to build on in a sustainable way.

It is now also appropriate and desirable to provide evaluations of alternative paradigms and technologies, that can inform TREs choosing between, for example, the principles-based manual approach supported by SACRO, the fully automated, strict rules-based approach implemented within DataShield, and various toolkits developed in the Machine Learning (ML) field which are not intended for OSDC but may have some use.

Recognising the significance of the topics, DARE identified '*Technology evaluation and comparisons in … Automated output checking for TREs (including evaluation of AI model outputs)*' as one of its priority areas for community groups.

This proposal is timely given:

- The level of interest in the SACRO tools arising from presentations at HDR, DARE, UK TRE Community and other national and international meetings
- The rising numbers and urgency of 'referrals' from TREs who have ML models that they need to risk-assess, but do not feel equipped to do.
- The overall DARE vision of inter-connected federated analytics wherein the need for disclosure control will escalate significantly and hence need to be supported by thoroughly tested and 'community-approved' automated tools for OSDC.

The challenges this group will address are:

1. Establishing a user community to:
    a. continue ongoing maintenance and development of the tools arising from the SACRO project
    b. embrace and evaluate other tools both extant and as they may arise; and
    c. establish protocols and mechanisms for evaluating different technologies related to OSDC.
2. Identifying and addressing blocks to impact- such as community created and delivered training for both researchers and TREs, and assurance for IT manager and governance groups at TREs.
3. Enabling TREs to support the creation and egress of machine learning models trained on confidential data. This long-term challenge has multiple facets which need to be addressed simultaneously:
    a. on-going development and maintenance of risk assessment / minimisation tools (for researchers and TREs)

   b. creating the *right* training resources for both TREs and researchers
   c. addressing the skills-gap at TREs around ML model privacy risks
   d. establishing a pool of experts and means of engaging them, that TREs can call on for help in assessing outputs they don't feel confident doing 'in-house' particularly the more complex ML cases.

4. Ensuring that the OSDC community, and any resources it develops, are informed by other community interest groups and developments, and will meet the future demands of federation and distributed data governance and OSDC.

Objectives

This community group primarily addresses the DARE theme "Capability and Capacity", however the work involved necessarily also impacts on, and will be informed by other work on "Demonstrating Trustworthiness", "Data and Discovery", and 'Core Federation Services".

It differs from other groups in the area through its focus on 'Safe Outputs', since without the capability to provide this assurance a key part of the 'Five-Safes' disappears, and DARE has recognised that manual checking currently presents a bottleneck. Moreover, studies before and during the DARE 'GRAIMatter' project established that many (possibly all) TREs do not feel equipped to assess the disclosure risk of machine learning models trained on sensitive data, and will need expertise they can draw on to provide support when they lack the technical understanding in-house.

The community group will initially work on four parallel threads, all underpinned by the Terms of Reference and the establishment of governance mechanisms for the open-source repositories. Naturally there will be overlap between involvement in these threads, and in the longer term we expect that new foci may emerge, and others diminish in priority.

Note that although we use SACRO for brevity, this should be taken to include a range of other tools for (semi) automated OSDC both extant (e.g. DataShield, ML-Privacy-Meter) and future developments.

Focus 1: Conceptual development and guidelines:

- **Rationale**: SACRO substantially changed perspective on Output Statistical Disclosure Control (OSDC) through creation of a framework and taxonomy and we now need to revise other materials and get community agreement
- **Activities**: (i) Workshop to: review and adopt material; identify need for further material e.g., alignment with SDAP manual (currently under revision); identify areas and priorities for future development and collaborations.
- **Outcomes**: Revised perspective on OSDC; established expert group to take forward future changes & roadmap

Focus 2: Removing barriers to adoption by researchers:

- **Rationale**: SACRO's testing has confirmed that the principle works. However, naturally adoption by users is currently very limited (although Eurostat report a growing uptake of the predecessor *acro* Stata tool). Feedback from the community is *where possible* to encourage, rather than force researchers to use specific toolkits (as per DataShield and tools at Stats Canada). Therefore, what is needed is community-led development of resources to enable and motivate researchers, such as training, videos, example code, enhanced help documentation, that can be accessed both *prior* to a 'TRE research session' and *during* the session (when external access is more limited)

- **Activities:** (i) Series of on-line and in-person workshops to design and approve resources to be developed such as user-centred training materials. (ii) Establishment of 'community researcher mentors' who can provide on-going support e.g., through regular advertised 'drop-in help sessions' and a central email support service.
- **Outcomes:** Roadmap for user adoption; resources; sustainable on-going researcher support network.

Focus 3: Enabling adoption by TREs

- **Rationale:** The TREs involved in the SACRO project all tested SACRO but achieved different levels of deployment within their secure environments, mainly due to separate infrastructure changes. We now need a sustainable network of support for TREs in: making a case for deploying automated checking (e.g., around IT governance/ software risk analysis); agreeing methodologies and tools to support evaluation (e.g., tools for 'reverse engineering' previous researcher code into the 'SACRO' framework to permit side-by-side comparison); on-going development of TRE-facing training materials; (iv) adding capability.
- **Activities:** (i) Monthly on-line / in-person/hybrid workshops to design and approve resources such as installation guides, governance protocols for code repositories, and mechanisms for community prioritisation of 'wish-lists'. (ii) Weekly on-line 'drop-in' help sessions for TREs at different stages of deployment and evaluation. (iii) Identification of `champions' amongst TRE community members who can provide peer-mentoring, and mechanisms for building this out sustainably. (iv) Outreach to other (DARE) community groups, especially regarding federated analytics.
- **Outcomes:** White paper comparing different technologies for automated OSDC. Governance structures for repositories. Range of materials and mentoring support for new organisations. Ongoing support and development of SACRO code repositories.

Focus 4: Risk assessment of Machine Learning models

- **Rationale**: SACRO, and prior to that GRAIMatter has established a range of guidelines, and mechanisms for automatically assessing disclosure risk of trained ML models according to several different metrics. However, this is a rapidly moving field, and conceptual still gaps exist between the ways that 'traditional OSDC and ML-privacy research consider risk, which SACRO has only partially been able to address. We need to establish of a community of expertise in interpreting ML risk metrics and explaining them to researchers and governance teams. This goes some way to addressing the skills gap identified during GRAIMatter – that it is probably not realistic to expect all TREs to maintain this expertise 'in-house'.
- **Activities:** (i) Establishing sustainable series of workshops around 'ML privacy risk in the context of TREs'. (ii) Establishment of a 'Community of Expertise' – a pool of experienced people and archive of experience around assessing specific ML models. (iii) Ongoing development and support for the AI-SDC code toolkit for ML risk assessment - to include new forms of attack as the field develops.
- **Outcomes:** Sustainable community of expertise enabling: closer alignment of OSDC and developments in risk assessment from ML researchers; on-going support and extension of risk-analysis toolsets; mechanisms for providing practical advice and support to researchers and TREs.

Outcomes: The initial foci and their proposed outcomes are listed above and may be summarised as:

- Alignment of conceptual framework and taxonomy of outputs with current and future training resources developed elsewhere.
- Establishing sustainable processes for the community-led design and implementation of resources to address methodological, 'practical' and training needs for the evaluation, deployment and increasing uptake of automated methods for OSDC.

- Establishing a ‘Community of Expertise around the OSDC of Machine Learning models including: training and tool resources for TREs and researchers, a living archive of best practice, and a pool of people who can be drawn on to provide decision support for TREs around ML models.

Participation/Collaboration: The community will initially include people from the following groups:

- Project leaders from the GRAIMAtter, SACRO and DataShield projects.
- TRE’s, and TRE-hosting organisations such as SAIL, keen to evaluate (semi) automated tools for OSDC for ‘traditional’ outputs and deploy them to address capacity and allow their staff to focus on more challenging cases.
- TREs keen to establish working mechanisms allowing them to support the creation and release of ML models trained on the data they hold.
- Computer scientists interested in the development of such tools and resources both for individual TREs, and to support federated analytics.
- AI/ML researchers interested in theoretical concepts and algorithms for assessing and quantifying the privacy disclosure risk of trained models *in the context of TREs*.
- Researchers, practitioners, and training-providers involved in the wider development of the concept of disclosure risk and what constitutes ‘Safe Outputs’.
- Public engagement practitioners focussing on public perceptions of trustworthiness and the ‘balancing the risk’ between privacy leakage and potential public good of research findings.

As the community becomes established and increases its outreach, we expect to widen the representation to include a range of organisations ranging from charities to National Statistics Institutes.

Mechanism:

The group will meet monthly, in a mixture of face-to-face, hybrid and online meetings. Each focus stream will self-organise under lead of a co-chair and may meet more often. They will organise targeted meetings with a wider audience- for example, a workshop in Liverpool hosted by DataShield to agree methodologies for evaluating and comparing approaches, and producing a white-paper. The Chair and co-chairs will meet fortnightly to review progress and balance effort between focus streams. As this is intended to be a living community, any member may propose new activities of focus, or changes to existing foci, for discussion at the monthly meetings.

The timelines of the call are short, and some people’s calendars can be fixed well in advance, limiting their ability to devote time to produce resources. Therefore, to create and sustain momentum for the community, we are specifically asking for support for a research software technician during this five-month phase to implement (prototypes of) various artefacts (resources, code etc.) designed by the members. That person, in combination with the chair and co-chairs will also host the ‘drop-in’ support sessions for researchers and TREs which will start to operate weekly from November 2023.

Potential members: [Including a minimum of two proposed chairs and all members who have expressed interest]

The people listed in the following table have confirmed via email. We have had positive meetings and verbal agreement but not official email confirmation from a range of other UK organisations such as OurFutureHealth. At the time of writing, we are currently presenting a number of papers about SACRO at the biannual international UNECE workshop on Statistical Disclosure Control. Following those presentations and subsequent discussions we have had verbal expressions of interest in joining the Community Group from the representatives of the following organisations, who have not had time to negotiate the bureaucracy of formal email approval: Eurostat (Aleksandra Bujnowska) , Stats-Netherlands (Peter-Paul de Wolff), Bundesbank, and the national Banks of Spain

(Eugenia Koblents), Italy (Giuseppe Bruno) and Bundesbank(several people interested, representation to be decided).

FIRST NAME	LAST NAME	EMAIL	(Co-)Chair / Member
Jim	Smith	James.smith@uwe.ac.uk	Chair
Jackie	Caldwell	Jackie.Caldwell3@phs.scot	Co-chair
Felix	Ritchie	Felix.Ritchie@uwe.ac.uk	Co-chair
Simon	Rogers	Simon.rogers@nhs.scot	Co-Chair
Amy	Tilbrook	Amy.tilbrook@ed.ac.uk	Member
Elizabeth	Green	Elizabeth7.Green@uwe.ac.uk	Member
Pete	Stokes	pete.stokes@phc.ox.ac.uk / pete.stokes@thedatalab.org	Member
Ben	Butler-cole	benjamin.butler-cole@phc.ox.ac.uk	Member
Christian	Cole	c.cole@dundee.ac.uk	Member
James	Liley	james.liley@durham.ac.uk	Member
Kate	O'Sullivan	katherine.osullivan@abdn.ac.uk	Member
Laura	Ennis	laura.ennis@phs.scot	Member
Alba	Crespi-Boixander	acrespi001@dundee.ac.uk	Member
Richard	Preen	Richard2.Preen@uwe.ac.uk	Member
Maha	Albashir	Maha.Albashir@uwe.ac.uk	Member
Susan	Kruger	SKrueger001@dundee.ac.uk	Member
Emily	Jefferson	Emily.Jefferson@hdruk.ac.uk	Member
Layla	Robinson	layla.robinson@researchdata.scot	Member
Katie	Oldfield	katie.oldfield@researchdata.scot	Member
Becca	Wilson	becca.wilson@liverpool.ac.uk (Data Shield)	Co-chair
Margaret	Levenstein	maggiel@umich.edu (ICPSR)	Member
Simon	Parker	simon.parker@dkfz-heidelberg.de (GHPH)	Member
Deborah	Wiltshire	Deborah.Wiltshire@gesis.org (GESIS)	Member
Chris	Dibben	Chris.Dibben@ed.ac.uk (SLS/SCADR)	Member
Andrew	Boyd	A.W.Boyd@bristol.ac.uk (UKLLC)	Member
Emma	Squires	emma@chi.swan.ac.uk (SAIL/Dementia Platforms UK)	Member
Steve	Harris	steve.harris@ucl.ac.uk (FlowEHR)	Member

** Note, please do not hesitate to point out gaps in the current DARE UK set of strategic themes and/or recommendations that the programme should consider as it continues to evolve these. Community feedback and input is welcomed.*