PDF Download
3748699.3749783.pdf
15 January 2026
Total Citations: 0
Total Downloads: 117

Published: 03 September 2025

Citation in BibTeX format

GoodIT '25: International Conference on Information Technology for Social Good
September 3 - 5, 2025
Antwerp, Belgium

Conference Sponsors:
SIGCAS

RESEARCH-ARTICLE

Sustainable Urban Mobility: Co-Designing a Responsible AI Recommender System

ALINA PATELLI, Aston University, Birmingham, West Midlands, U.K.

ANIKÓ EKÁRT, Aston University, Birmingham, West Midlands, U.K.

MARGARITA CHLI, Aston University, Birmingham, West Midlands, U.K.

LAVINIA EUGENIA FERARIU, "Gheorghe Asachi" Technical University, Iasi, Iasi, Romania

JOHN REGO HAMILTON, Aston University, Birmingham, West Midlands, U.K.

MERCY KANYI, Aston University, Birmingham, West Midlands, U.K.

View all

Open Access Support provided by:

Aston University

Ontario Tech University

"Gheorghe Asachi" Technical University

Glasgow Caledonian University

GoodIT '25: Proceedings of the 2025 International Conference on Information Technology for Social Good (September 2025)
https://doi.org/10.1145/3748699.3749783
ISBN: 9798400720895

Sustainable Urban Mobility: Co-Designing a Responsible AI Recommender System

Alina Patelli
Aston University
Birmingham, United Kingdom
a.patelli2@aston.ac.uk

Anikó Ekárt
Aston University
Birmingham, United Kingdom
a.ekart@aston.ac.uk

Maria Chli
Aston University
Birmingham, United Kingdom
m.chli@aston.ac.uk

Lavinia-Eugenia Ferariu
Technical University of Iasi
Iasi, Romania
lferaru@ac.tuiasi.ro

John Hamilton
Aston University
Birmingham, United Kingdom
j.connorregohamilton1@aston.ac.uk

Mercy Kanyi
Aston University
Birmingham, United Kingdom
m.kanyi@aston.ac.uk

Richard Lee
Aston University
Birmingham, United Kingdom
r.lee6@aston.ac.uk

Peter Lewis
Ontario Tech University
Birmingham, United Kingdom
Peter.Lewis@ontariotechu.ca

Joanna Lumsden
Glasgow Caledonian University
Glasgow, United Kingdom
Joanna.Lumsden@gcu.ac.uk

Stephen Owen
Aston University
Birmingham, United Kingdom
s.owen@aston.ac.uk

Abstract

Responsible AI is a tech driver of sustainable economic growth that protects democratic liberties. The systematic design, implementation, and deployment of AI for good are demanding tasks, given the diversity of those impacted. Engaging a representative sample of AI's heterogeneous user base to gauge the benefits it expects requires innovative participatory activities interspersed throughout the stages of the AI development process. Translating stakeholder input from the jargon-free vocabulary in which it is collected to coherent, comprehensive, industry-standard artefacts that experts can use to build responsible AI in practice is also challenging. Developing a robust assessment framework with objective metrics for evaluating intelligent tech adds to the overall difficulty. We capture these aspects in seven key challenges which we address by proposing a novel, systematic, participatory approach to co-designing and co-assessing responsible AI. We apply the approach to architect an AI recommender system that supports transport authorities, industry, policymakers, and the public with their urban mobility decisions. Throughout three workshops, representatives from the four stakeholder categories worked with domain experts to co-develop *a system blueprint* featuring complementary tech from across the AI spectrum and *an evaluation framework* with robust blueprint assessment metrics. This paper presents the two artefacts alongside a detailed account of the innovative workshop activities leading to their co-creation.

GoodIT '25, Antwerp, Belgium

ACM ISBN 979-8-4007-2089-5/25/09
https://doi.org/10.1145/3748699.3749783

CCS Concepts

• **Human-centered computing → Collaborative and social computing**; • **Information systems → Decision support systems**; • **Computing methodologies → Model development and analysis**; *Planning and scheduling*.

Keywords

Responsible AI for social good, Participatory design, AI recommender systems

ACM Reference Format:
Alina Patelli, Anikó Ekárt, Maria Chli, Lavinia-Eugenia Ferariu, John Hamilton, Mercy Kanyi, Richard Lee, Peter Lewis, Joanna Lumsden, and Stephen Owen. 2025. Sustainable Urban Mobility: Co-Designing a Responsible AI Recommender System. In *International Conference on Information Technology for Social Good (GoodIT '25), September 03–05, 2025, Antwerp, Belgium.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3748699.3749783

1 Introduction

Balancing the drive to build ever more powerful AI that supports economic growth and improves the standard of living with incentives to make public-facing intelligent tech as safe and inclusive as possible is a strategic target of most governments and organisations. To realise that goal, it is imperative to develop and consistently apply systematic approaches to building value-adding responsible AI, ensuring that all stakeholders (end users, tech experts, policymakers, etc.) are involved throughout design, development, and assessment, and that their requirements are collated in coherent formats well-suited to informing AI tool production.

We address this need by proposing an innovative, principled participatory research process that prescribes the design and evaluation of AI recommender systems which are trustworthy, ethical, explainable, sustainable, equitable and inclusive, in one word, responsible. Without restricting its generality, we apply the process to the intelligent urban mobility problem domain to prove its effectiveness in the real world. The case study consists of three residential workshops that feature a carefully constructed sequence of stakeholder-centric activities designed to optimally engage heterogeneous audiences, creating ideal conditions for collaborative and impactful work. The work's aim was to architect an Intelligent urban Mobility (iU&Mi) digital platform that issues actionable, accurate, and explainable recommendations for (1) the public making travel arrangements, (2) the city's transport authorities managing road infrastructure, (3) the service industry operating vehicle fleets, and (4) the policymakers in transport governance. The workshops elicited stakeholder requirements, i.e., iU&Mi's real-world impact from the viewpoints of its target audiences, led to co-designing an architecture underpinning iU&Mi's implementation, and enabled co-evaluating, with objective metrics, the architecture's potential to yield a functioning, fit-for-purpose recommender system. To make the three workshops successfull, the authors faced seven key chalenges.

- C1 In order for the finished recommender system to meaningfully benefit all stakeholder categories, iU&Mi's **heterogeneous target audience** needed to be adequately represented amongst a carefully calibrated mix of workshop participants.
- C2 All urban mobility needs had to be elicited via **engaging, meaningful activities**, systematically structured to channel the stakeholders' creative and critical reflection efforts towards producing insightful outputs, genuinely helpful to system developers when producing the finished product.
- C3 Effective residential activities rely on a **carefully calibrated number of participants**, large enough to provide a diversity of views leading to insightful results and small enough to give everyone a chance to contribute.
- C4 **Continued stakeholder involvement** throughout the scoping, design and assessment of the recommender system architecture is a cornerstone of participatory research.
- C5 Workshop tasks needed to be set in an **accessible vocabulary** that is jargon-free and straightforward to translate to the standard formats and templates systems engineers use in their development work.
- C6 Powerful recommender systems exploit the strengths and mitigate the weaknesses of several complementary AI components, from across the evolutionary and deep learning spectrum, expertly selected and assembled into a **coherent design**, one that can be implemented in practice.
- C7 The recommender system's architecture needed to be systematically and comprehensively evaluated throughout the various iterations of the design process via a robust set of metrics forming an **assessment framework** co-defined by the target audience (requirement owners) and the experts (architecture authors).

The workshop series successfully addressed these challenges and produced a coherent and comprehensive requirements document, capturing the diverse range of iU&Mi features and non-functional characteristics that the stakeholders deemed beneficial. It also led to the creation of an innovative template illustrating iU&Mi's conceptual architecture, one that coherently combines expertise from across the AI spectrum (evolutionary, deep learning, AI ethics, etc.) and conforms to the principles of system engineering, human-centric computing, and data science. The stakeholders and experts also developed a systematic assessment framework to objectively measure the template's quality.

In this paper, we present the iU&Mi case study (section 3) and its outputs (the requirements document, the recommender system architectural template, and the evaluation framework) through the lens of responsible AI. The focus of the paper is not on the algorithmic achievements, but on the participatory, stakeholder-centric way these were developed. Section 5 summarises the lessons learnt from applying the proposed participatory research approach to the intelligent mobility case study.

2 State of the Art

AI is ubiquitous in the public domain where recommender systems permeate people's digital lives. On social media platforms, likes and posts are analysed to suggest relevant content. Wearable devices monitor biometrics and mobility patterns to recommend behavioural changes that help users reach fitness goals or health targets more easily. Intelligent algorithms process historical data to inform medical diagnoses, court sentences, and credit score calculations. The escalating rate at which AI is being adopted across society has prompted growing concerns over the safety of intelligent tools [19, 30], with governments and industry across the world championing legislative, operational, technological, and social initiatives to regulate the design, development, deployment, and long-term use of AI. The UK AI Safety Institute, the EU AI Act and similar initiatives in the US, China, and several other major tech-intensive economies are only a few examples of such efforts. Researchers are also active in this field, showing a growing interest in the impact of AI on sociotechnical systems [3, 21], contributing to an expanding body of work on the systematic evaluation of AI trustworthiness [7, 18, 38], and collaborating with a thriving community of scholars and practitioners to promote digital fairness, accountability, and transparency [6, 13, 17]. Big Tech is following suit, with responsible AI research and development gaining momentum at Google [4, 11, 37], Meta [1, 2, 14], major streaming platforms [24, 32, 35], and Microsoft [10, 27].

The interplay between the creators, vendors, legislators and consumers of responsible AI, together with the implications on society as a whole are thoroughly investigated in techUK's most recent white paper on the AI profession [33]. The report stresses the importance of approaching the development of intelligent algorithms in a methodical, participatory fashion where practitioners, consumers and regulators collaborate to embed ethical principles throughout the scoping, design and evaluation of AI tools for the benefit of all.

Whilst this acute need for systematic ways of co-creating responsible AI has been addressed from relevant yet disjointed perspectives such as explainability [31], environmental sustainability [22], and impact on cognition [12, 20], integrated approaches that tackle the problem from a holistic standpoint are relatively rare.

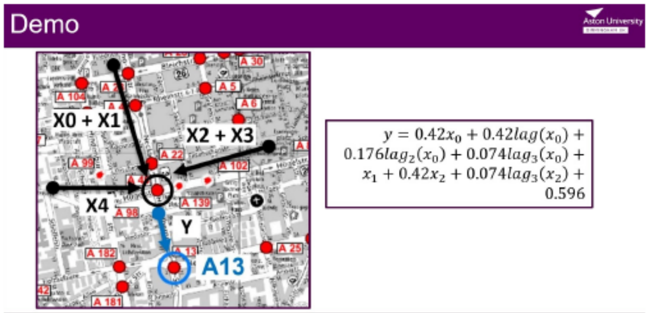
Several participatory research strategies, frameworks and best practice attempt to bridge that gap. Action research [9], asset-based community development [25], community-engaged research [16], decolonising and emancipatory methodologies [34], pragmatic action research [15] and user-centred design [23] are established ways routinely employed by scientists and stakeholders when co-creating solutions with a transformative social impact. The next step, not yet taken, is to adjust and combine components of these various frameworks and strategies to construct a bespoke methodology, and refine it by adding original elements. This is the rationale behind our principled participatory approach to co-designing responsible AI systems, a timely contribution towards a strategic goal of governments, organisations and urban communities everywhere.

3 The iU&Mi Case Study

Our systematic approach to developing responsible AI recommender systems was applied to a concrete participatory research study consisting of three structured workshops where a diverse group of stakeholders engaged in innovative activities to express their urban mobility requirements (3.1), design the architecture of a city traffic recommender system called iU&Mi (3.2), and produce an evaluation framework with concrete metrics for measuring the architecture's quality (3.3). We present the way the workshops were planned and run, and the outputs of the workshop activities. The discussion around each workshop activity, apart from ice-breakers, will cover a brief description of the task, the specific challenges it addresses (out of the seven in the introduction), and how it fits within the goal, reality, opportunities, work, evaluation (GROWe) process (adapted from [29] by adding an evaluation stage).

3.1 Requirements elicitation workshop

Participants Attendees included ten members of the general public (commuters, fitness professionals, motorists, pedestrians, public transport users), two representatives of urban traffic administration (one senior departmental lead within Transport for West Midlands and an infrastructure and sustainable engineering expert), three industry stakeholders (from British Telecom and a Birmingham-based family-owned bakery), and two policy professionals (one key contributor to West Midlands transport governance and one specialist in AI tools for public engagement with Government).
Ice breaker Participants stated their professional and personal drivers, interest in intelligent urban traffic recommendation systems, and expectations from the day. This addresses challenge C1 and is a precursor to tackling challenge C2.
Demo The facilitators demonstrated GENTLER (GENetic programming with Transfer LEarning and Randomisation) [28], the intelligent traffic prediction algorithm meant to power iU&Mi, on the simple junction topology illustrated in Fig. 1. The algorithm's output is a straightforward mathematical model equating outflow traffic to a combination of the three inflow volumes. The facilitators showed the audience how each term of the equation reveals valuable insight about real world traffic: likely congestions, potential delays at stop lights and lane occupancy. This engaging, accessible demo about a relatable situation routinely experienced in day-to-day traffic was more meaningful to the mix of specialist and non-specialist participants than going through conventional algorithm specification documentation; this addresses challenges C2 and C5. Within the GROWe approach, the demo maps against the goal and reality-setting phase, allowing the facilitators to clearly define the overarching aim of the workshop series, namely embedding GENTLER, a tool requiring technical know-how to run in order to generate models that take specialist expertise to interpret and use, into a versatile recommendation system, providing value to all stakeholders, irrespective of their level of digital skill.



Demo

$$y = 0.42x_0 + 0.42lag(x_0) + 0.176lag_2(x_0) + 0.074lag_3(x_0) + x_1 + 0.42x_2 + 0.074lag_3(x_2) + 0.596$$

Figure 1: Map: junction in Darmstadt, Germany; black arrows: inflow traffic; blue arrow: outflow traffic. Equation: GENTLER-generated traffic model of inflows combining into outflow. The model reveals traffic patterns that GENTLER learns from historical sensor data (retrieved from https://datenplattform.darmstadt.de/verkehr/apps/opendata/#/): (1) X0 and X2 vehicles are stalled before exiting the junction (as indicated by the lags in the model); (2) X1 traffic has the greatest influence on outflow (coefficient is 1); and (3) X4 traffic does not impact the outflow at all. Decision makers can draw valuable insight from these patterns to: (1) lighten congestion by reprogramming the junction's stop lights, widening the lanes or adding overflow arms to distribute inflow; (2) consider other topologies where all arms are relevant; and (3) repurpose X4 as a pedestrian or bicycle lane.

Requirements synthesis In the TREQ (TREe of reQuirements) activity (adapted from [29]) the participants filled in tree-shaped diagrams with their transport-related problems (trunk), their causes (roots), proposed solutions (branches) and their envisaged positive and detrimental impact (fruit). This engaging, visual, problem-solving exercise (Fig. 2) is more structured than an open ended discussion, where contributions are less likely to coalesce into coherent outputs, yet less rigid than formal requirements elicitation (e.g., survey, interview, focus group, etc.) that may intimidate participants or fail to spark their enthusiasm. The original activity proposed in [29] was enhanced with three innovative mechanisms. Firstly, the participant groups switched work stations after completing their TREQ to provide feedback on their peers' trees. Groups then returned to their original stations and addressed peer comments. Finally, teams shared their final TREQ and reflect on its genesis. The enhanced TREQ activity promotes collaboration in a relatable, visual and inviting manner, free of technical jargon, addressing challenges C2 and C5. The exercise prompted participants to consider both positive and (unintended) negative impacts of proposed solutions (healthy and wormy fruit, respectively), provide and address peer feedback (work station switching), and share critical summaries of their work (final presentations). The professional execution of the activity, the structured collaborative work involved and the polished outcomes produced reassured stakeholders that their resources were invested in a worthwhile enterprise and not squandered on playing whimsical games. The work station switching mechanism achieved a trade-off between group size (small enough to allow everyone a chance to contribute) and inclusivity (output of one group is reviewed by peers to ensure all views are accounted for). This addresses challenge C3. The inclusivity afforded by switching and critical summary presentations eliminated most conflicts and redundancies between the groups' TREQs making it easier to translate the trees into one coherent requirements document, set in the standard format that software engineers can work with (a precursor to addressing challenge C6). Adding MoSCoW priorities [26] and evaluation metrics (challenge C7) were also simplified as a result. The TREQ activity fits within the opportunities phase of GROWe: it sets out ways to turn goals into reality.

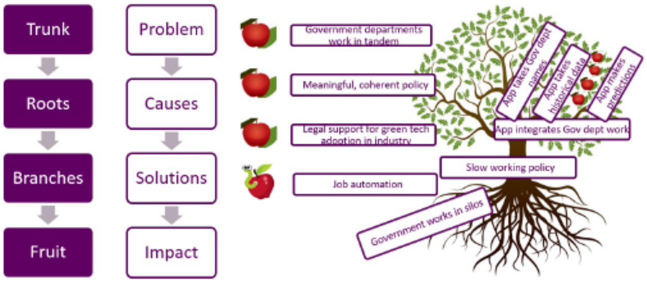

Trunk
Roots
Branches
Fruit
Problem
Causes
Solutions
Impact
Government departments work in tandem
Meaningful, coherent policy
Legal support for green tech adoption in industry
Job automation
App takes Gov dept names
App takes historical data
App makes predictions
App integrates Gov dept work
Slow working policy
Government works in silos

Figure 2: Requirements synthesis. The participant groups co-developed the generic TREQ outline (left) into concrete TREQ instances. The TREQ on the right led to requirements R5 and R10 (Fig. 3). The latter expanded the scope of the TREQ problem from national governance to the wider policymaking community (urban transport administrators, vehicle fleet managers and pedestrians planning day-to-day travel also tend to work in silos). Requirement R5 mitigates the TREQ's negative impact (next to the wormy apple).

Requirements assessment The participants applied the PESTLE technique to reflect on political, economic, social, technological and environmental aspects of their requirements' impacts (the fruit in Fig. 2). This scenario-based method for strategic decision making enabled the participants to derive coherent, realistic metrics for each requirement, as opposed to a 'wish list' of nice-to-haves. Combining the structure afforded by PESTLE with the relatable, intuitive TREQ (C5) allowed for a multi-faceted impact analysis (C2) at a low risk of intimidating or overwhelming the non-technical audience (more likely to happen when PESTLE is applied in isolation). Thus, the participants seamlessly transitioned to a mindset conducive to systematic assessment, making it easier for the facilitators to explain the principles of quantitative and qualitative analysis without relying on textbook definitions and risking alienating the audience. Guided by these gently introduced theoretical concepts, the participant groups engaged in informed collaborative work and formulated metrics largely free of conflicts and redundancies, therefore easier to collate in a coherent requirements evaluation framework (C7) that experts could directly refer to in the application design stage. As before, work station swaps and final group presentations ensured that substantial and varied peer feedback was provided to all and reflected upon constructively (C1, C3). Within GROWe, requirements assessment is the object of the evaluation stage.

Output The requirements document (excerpt in Fig. 3 shows three out of the eighteen entries) translates the TREQs put forward in the synthesis activity to the standard format used in traditional requirements engineering (that lists labelled requirements against associated metrics) enriched with annotations derived from the stakeholders' collaborative work.

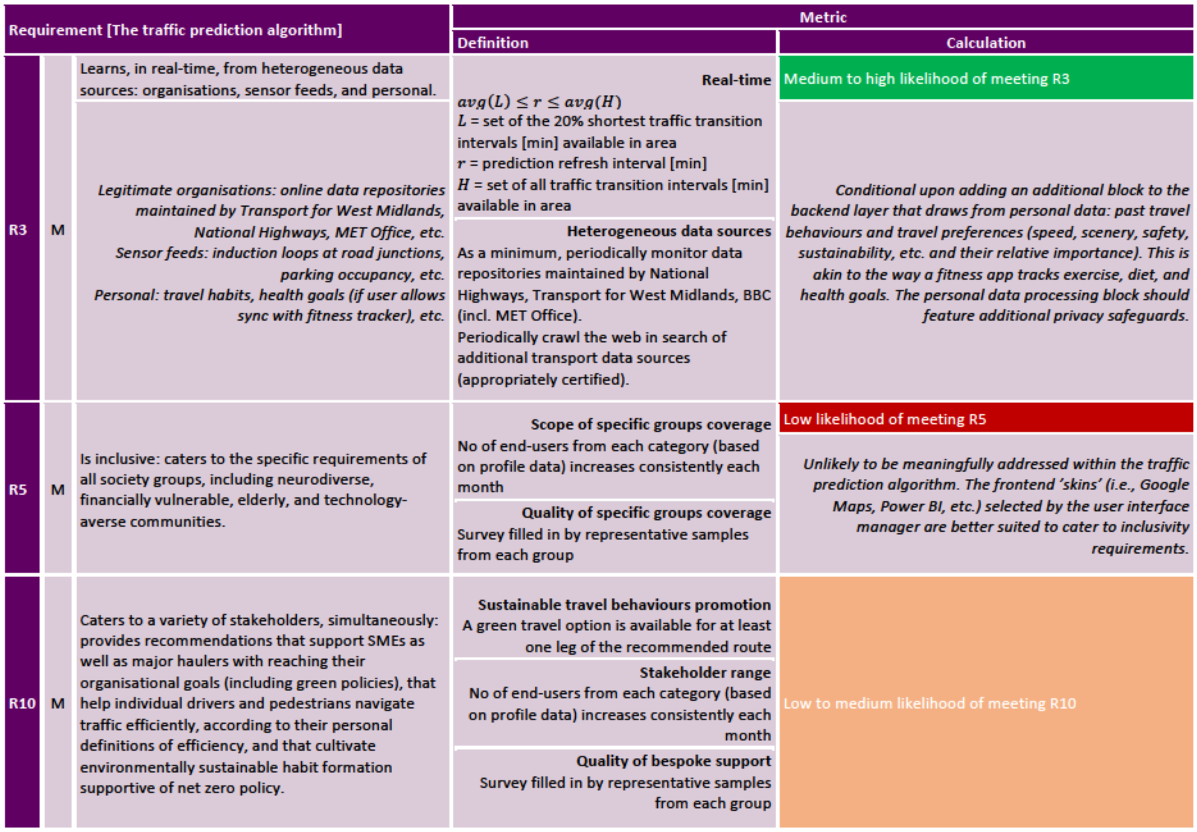
3.2 Architecture design workshop

Participants The experts in attendance comprised four specialists in evolutionary AI and four from complementary fields: deep learning, agent-based modelling and optimisation, and self-adapting, self-organising socio-technical systems. They were joined by four colleagues with expertise in trustworthy AI, human-computer interaction, data science, and software engineering.

Introduction The ice breaker followed the same structure as in the previous workshop. This consistency was designed into all sessions to prepare the diverse participants for efficient collaborative work by observing familiar practices (challenges C1 and C2).

Requirements fine-tuning The experts contributed their specialist perspectives to constructively analyse stakeholder requirements and translate them into specific, measurable, achievable, relevant, and time-bound objectives. A standard tool in professional software project management, the SMART framework [5] enabled the experts to eliminate the subtle ambiguities, redundancies and conflicts in the requirements document, getting it ready for the iU&Mi implementation phase. The requirement priorities and metrics proposed by the stakeholders were also revised (challenge C7); for instance, the data sources mentioned in R3 in Fig. 3 were further discussed to clarify the difference between data owners (i.e., the organisations collating and publishing road sensor readings and user devices monitoring mobility patterns) and data types: ordinary (vehicle counts, etc.), realtime (accidents, weather, maintenance works) and contextual (metadata such as sensor location, time of day and year, co-occurrence with major sporting events or festivals, proximity to landmarks, etc.). The fine-tuning of the stakeholder requirements by translating their initial versions into SMART objectives and then turning those back into (final) requirements was a collaborative, interdisciplinary exercise underpinned by a group work methodology similar to that applied in the requirements elicitation workshop: experts were assigned to groups that were diverse, i.e., included representatives from most disciplines (challenge C1), and compact, such that all members had opportunities to contribute to the discussion (challenge C3). As before, the work station swaps and group presentations enabled the provision of and reflection on peer feedback in the lead up to communicating the revised group work outputs to the entire audience and collating those to form the finished requirements document. In the GROWe framework, requirements fine-tuning is part of the work stage.

Architecture design This task presented three key difficulties.



Requirement [The traffic prediction algorithm]			Metric	
			Definition	Calculation
R3	M	Learns, in real-time, from heterogeneous data sources: organisations, sensor feeds, and personal. *Legitimate organisations: online data repositories maintained by Transport for West Midlands, National Highways, MET Office, etc.* *Sensor feeds: induction loops at road junctions, parking occupancy, etc.* *Personal: travel habits, health goals (if user allows sync with fitness tracker), etc.*	**Real-time** $avg(L) \leq r \leq avg(H)$ L = set of the 20% shortest traffic transition intervals [min] available in area r = prediction refresh interval [min] H = set of all traffic transition intervals [min] available in area **Heterogeneous data sources** As a minimum, periodically monitor data repositories maintained by National Highways, Transport for West Midlands, BBC (incl. MET Office). Periodically crawl the web in search of additional transport data sources (appropriately certified).	🟩 Medium to high likelihood of meeting R3 *Conditional upon adding an additional block to the backend layer that draws from personal data: past travel behaviours and travel preferences (speed, scenery, safety, sustainability, etc. and their relative importance). This is akin to the way a fitness app tracks exercise, diet, and health goals. The personal data processing block should feature additional privacy safeguards.*
R5	M	Is inclusive: caters to the specific requirements of all society groups, including neurodiverse, financially vulnerable, elderly, and technology-averse communities.	**Scope of specific groups coverage** No of end-users from each category (based on profile data) increases consistently each month **Quality of specific groups coverage** Survey filled in by representative samples from each group	🟥 Low likelihood of meeting R5 *Unlikely to be meaningfully addressed within the traffic prediction algorithm. The frontend 'skins' (i.e., Google Maps, Power BI, etc.) selected by the user interface manager are better suited to cater to inclusivity requirements.*
R10	M	Caters to a variety of stakeholders, simultaneously: provides recommendations that support SMEs as well as major haulers with reaching their organisational goals (including green policies), that help individual drivers and pedestrians navigate traffic efficiently, according to their personal definitions of efficiency, and that cultivate environmentally sustainable habit formation supportive of net zero policy.	**Sustainable travel behaviours promotion** A green travel option is available for at least one leg of the recommended route **Stakeholder range** No of end-users from each category (based on profile data) increases consistently each month **Quality of bespoke support** Survey filled in by representative samples from each group	🟧 Low to medium likelihood of meeting R10

Figure 3: A sample from the final requirements document. The three requirements shown illustrate the way stakeholders formally assessed the iU&Mi architecture's capacity to yield a finished product that delivers on their needs. The likelihood of that outcome is high (green), medium (amber), or low (red). Calculations are annotated with justifications

Frontend interface Although commonalities exist, public administrators, industry representatives, members of the public and, respectively, policymakers require the app to deliver recommendations specifically tailored to their contexts, in interpretable formats directly applicable to their bespoke circumstances (challenge C1). This calls for several iU&Mi frontend variants, following a consistent yet flexible design that can adjust the app's output according to the profile of the current user. Providing the recommender system with these interchangeable 'skins' demands technical insight as well as human-computer interaction expertise.

Core functionality Actionable and reliable recommendations are based on robust traffic predictions drawing from historical and current data collected from the area of interest or transferred from a topologically similar one. This level of predictive quality requires the cumulated functionality of several complementary AI components expertly combined to form a working software platform. Specialists from across the AI spectrum (evolutionary computation, deep learning, responsibility-driven AI design) worked with system engineers to select suitable components and integrate them into a holistic architecture, where the weaknesses of individual parts (e.g., the opaqueness of deep learning models) are mitigated by the strengths of others (the intrinsic explainability of evolutionary algorithms) (challenge C6).

Data processing The recommender's data layer is responsible for the acquisition, storage, and low-level processing of heterogeneous urban mobility inputs: pedestrian and motorised traffic sensor readings, personal travel preferences, imports from third party apps on user devices, high-level policy and city-wide transport administration strategy, etc. These are essential sources of knowledge for training robust models that can yield reliable predictions. The platform's data management element thus needed a hybrid architecture blending together established data analytics modules (relational records of sensor readings labelled CDF in Fig. 5) with bespoke components creatively designed to meet stakeholder requirements (the persistent models in Fig. 5) (challenge C6).

To tackle these design complexities, the experts started with the industry-standard three-layer app construction: a frontend, a middle tier hosting the algorithms underpinning the system's functionality, and a data processing backend. While the familiarity of this established architecture instilled confidence, it also presented the risk of complacently following 'the trodden path' and unintentionally omitting potentially superior design alternatives. To prevent that, the activity featured a prompt to consider wildcard structural variants, should they emerge during group discussions. The wildcard prompt placed the experts in an innovative mindset, helping them creatively combine tailored frontend interface instances, evolutionary and deep learning elements in the app's middle tier, and specialised data processing back-end modules, all non-standard architectural elements that a traditional approach bound by the standard practices of mainstream software engineering would have made less likely to consider (challenge C2). Experts were put into three groups each assigned a layer of the architecture. The frontend group included human-computer interaction specialists, the one working on the middle tier gathered experts in systems engineering and AI algorithm developers, whilst the majority of the back-end team consisted of data scientists. AI ethicists and responsibility-driven design experts were appointed to each group. This carefully calibrated group membership (challenge C3) successfully prevented working in silos (i.e., drawing exclusively from specialist knowledge directly applicable to the assigned architecture layer) which would have most likely resulted in three disconnected sub-designs difficult to assemble into a coherent holistic architecture. The groups underwent two rounds of swaps, such that frontend design considerations could be infused into the middle tier and, respectively, back-end construction and vice versa (challenges C2 and C6). Each group conveyed the findings from their summative reflection on peer feedback during lightning presentations. Designing the iU&Mi architecture fits into the opportunities, work and evaluation stages of GROWe, as it illustrates a path (blueprint) to get from goals to reality, involves collaborative work to develop said blueprint with its practical implementation in mind, and is informed by the metrics proposed in the first workshop to measure its ability to deliver on stakeholder requirements.

Preliminary architecture evaluation Following the assembly of the three layers into a coherent architecture blueprint, the experts engaged in an evaluation activity to refine and calculate the stakeholder metrics. To ensure consistency, the activity was organised in the same fashion as previous ones: the experts worked in groups, switched stations to input into their colleagues' outputs, reflected on peer feedback and presented their conclusions in brief talks. The work produced during the preliminary blueprint evaluation activity contributed towards the work and evaluation phases of GROWe.

Output The design workshop resulted in a revised version of the requirements and associated metrics (Fig. 3). The iU&Mi architecture, the second artifact developed by the experts and further refined in the evaluation workshop, is illustrated in its standard UML format (Fig. 4) and in a visually engaging rendition (Fig. 5). The two are logically equivalent: the former is ready to use in the software development stage, whilst the latter was specifically created for the benefit of the non-technical stakeholders and to simplify the third workshop's evaluation tasks.

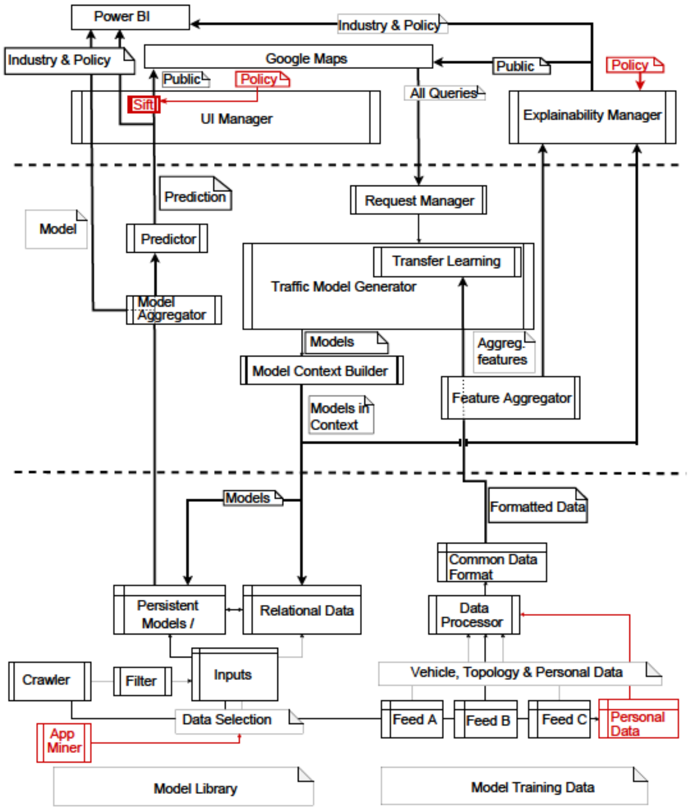


Figure 4: The iU&Mi architecture diagram in UML. Produced by the experts in the second workshop and revised by the stakeholders in the evaluation workshop (revisions in red). The frontend (above the top dotted line) dynamically changes the system 'skins' depending on current user profile: the general public views travel recommendations in the familiar Google Maps format, whilst policymakers, public administrators and industry receive insight in business-standard interfaces (PowerBI). All iU&Mi output is accompanied by a natural language explanation of how it was obtained. The middle tier (between the two dotted lines) assembles evolutionary AI (the Traffic Model Generator block) and deep learning components (the Feature Aggregator) to process, merge and enhance pre-existing traffic models and, if necessary, produce new ones. The predictive output of these models underpins the frontend recommendations. The backend (under the bottom dotted line) collects and aligns mixed traffic data from road sensors, user devices, etc. and trains the models. The historical library stores building blocks for future models, avoiding the need to train from scratch.

3.3 Evaluation workshop

Participants Industry representation was kept at the same level with new joiners from two transport consultancies and one community-centric immersive arts organisation. The urban transport administration group welcomed a senior strategy consultant for Midlands Connect, whilst policymakers were joined by a research governance expert with Innovate UK. Four new participants enriched the general public stakeholder group by contributing their day-to-day urban traffic experiences as well as expert insights into transport-relevant AI and cybersecurity. This slightly larger yet still manageably-sized mix of legacy participants and newcomers maintained the proportionality of representation across all stakeholder categories as well as the project's structural and thematic continuity, all the while injecting fresh perspectives into the audience. This trade-off proved conducive to efficient collaborative work leading to outputs which incorporated previous workshop deliverables as well as novel inputs.

Introduction Besides the opportunity to get re-acquainted with each other and meet new participants, the introduction also included a brief summary of project developments to date, refreshing the stakeholders' memory of the requirements elicitation workshop, updating them on the experts' progress and re-accustoming them to collaborative work patterns. This addresses challenges C1 and C2 and sets the scene for C4 and C7.

Architecture discussion The facilitators explained the traditional architecture diagram (Fig. 4) and its more visually engaging variant (Fig. 5). Whilst the latter increased the accessibility of an otherwise dry and intimidating software development artefact (challenges C2 and C5), the former served as concrete proof that stakeholder input was standardised to drive specialist work. This assuaged concerns over the intentional whimsy of the visually-engaging blueprint variant. The activity contributed to the opportunities and work phases of GROWe.

Architecture evaluation The participants discussed the expert-refined metrics and approved, altered or reverted their suggestions. This task completed the first iteration of the evaluation cycle, which started with the metric definitions proposed in the requirements elicitation workshop, continued with the experts' refinement of those metrics and their preliminary calculation, and concluded with the stakeholders' second round of edits (challenges C4 and C7). Once reviewing the blueprint's evaluation criteria was finalised, the stakeholders collaboratively calculated the revised metrics by answering the question 'What is your level of confidence that the digital tool(s) to be built from the proposed blueprint will achieve the metric associated to each requirement?'. Putting the evaluation task in the form of a question formulated in plain, accessible language (challenge C5) as opposed to a complex mathematical calculation streamlined the stakeholders' work, encouraging their engagement (challenge C2). For each requirement, the groups selected a likelihood of successful implementation (low, medium or high), with explanatory notes to justify their verdict and suggest future improvements. To continue addressing challenges C1, C2 and C3 consistently with previous practice, both exercises in the architecture evaluation activity followed the familiar structure: the stakeholders worked in teams with appropriate representation from all four categories, switched work stations to review and provide feedback on their colleagues' metrics and calculations, returned to their original stations to reflect on their peers' comments and presented their final outputs to the entire audience before the workshop's conclusion. The activity completed the initial iteration of GROWe's evaluation stage.

Output Besides performing metric calculations, the stakeholders reflected on their assessment and suggested ways to improve their success estimates, captured in the annotations shown in Fig. 3. The first change they suggested involved adding a filter to the frontend pipeline producing travel recommendations for the general public which are to be tailored in accordance with urban administrators' policy (e.g., even though driving may be the fastest option, iU&Mi would provide cycling and walking alternatives to promote the city council's green mobility strategy). This filtering would ensure that meeting user preferences is not at odds with the city's sustainable development policy and long-term planning. The second change introduced a backend training data stream supplying the core modelling and prediction algorithm with contextual knowledge on the users' historical mobility patterns, health targets and biometrics, etc., imported from other apps and wearable devices. This would increase the algorithm's sensitivity to personal circumstances, such as being a wheelchair user or a caretaker, leading to travel recommendations better suited to individual profiles. iU&Mi's inclusivity would thus improve. Both changes are shown in red in Fig. 4; their logical equivalents in the graphical version of the blueprint (Fig. 5) are rendered as an annotated funnel in the frontend and, respectively, a personal data graphic encased in a dotted wedge stemming from a magnifying glass in the backend.

4 Lessons Learnt and Preliminary Evaluation

We analyse the strengths and areas for improvement of the proposed participatory approach to co-designing and co-assessing responsible AI recommender systems, by reflecting on lessons learnt from the iU&Mi case study. The practical value of our core innovations is discussed in relation to the seven key challenges and in terms of the added benefits relative to established participatory co-design techniques [8, 29]. This affords a preliminary evaluation of our proposal, setting the foundation for the robust participatory research methodology the authors will develop and quantitatively assess.

Participant selection A diverse group of stakeholders was on site at all workshops (their testimonials in [36]). This successfully addressed C1, by reflecting the target audience's wide range of opinions, and C3, by creating work groups small enough to enable all voices to be heard yet sufficiently large to yield meaningful outputs. Each session was attended by a kernel of previous participants and received a controlled influx of newcomers ensuring a healthy mix of 'legacies' (continuity) and 'joiners' (fresh perspectives).

Collaborative tools Leveraging established tools and frameworks (MoSCoW, PESTLE, SMART, three-tier software architecture model) increased task professionalism. This engendered stakeholder trust in the process, encouraging them to confidently contribute their efforts towards meaningfully driving the apps' implementation, knowing that their energy investment was not squandered on whimsical games (C2). Moreover, the experts' pre-existing knowledge of established techniques afforded them the comfort of starting work from a familiar place. Far from stifling innovation, this made participants more at ease with proposing new ideas than they would have been if confronted with completely unexplored spaces. Enhancing established tools and models with original adjustments (GROWe, TREQ) to fit the workshops' agile working style established a creativity-prone environment, fostering ideation and lateral thinking. The experience taught us that the mix of established tools and innovative techniques could be fine-tuned to better prescribe the requirements definition task and reduce some of its ambiguities which, in this instance, have complicated the experts' design work.

Activity design The demo provided a practical showcase of the proposed tech on a simple relatable example, which grounded the requirements elicitation process without overly constricting it. The work station switches and summative presentations embedded into

activity design addressed C2 whilst accompanying the UML architecture diagram with a visually engaging equivalent met C5. These original task-underpinning mechanisms achieved a suitable trade-off between efficiency and creativity which enabled the experts to unpack the requirement owners' thinking and assumptions, identify and align stakeholder needs that were formulated at incongruous levels of abstraction, and build a cohesive app design (C6). New collaborative work techniques will be considered to refine the efficiency-creativity trade-off: role playing, story boarding, and other classic and emerging approaches are promising candidates.
Continuous evaluation Co-defining concrete metrics to form a robust framework for evaluating iU&Mi's design in terms of its ability to meet stakeholder requirements is key to responsible AI development. Involving participants early-on in the app assessment process and periodically seeking their input in subsequent stages (i.e., after the experts' revision of the initially proposed metrics and throughout implementation and deployment) successfully addressed C4 and C7. It also revealed an area for improvement: the three-tier quality metrics that the iU&Mi blueprint was assessed against (Fig. 3) afforded a coarse preliminary evaluation which could benefit from a more finely-grained grading system (perhaps numerical rather than nominal). This would warrant a more insightful view into the strengths and vulnerabilities of the proposed architecture, which would steer the app's implementation more efficiently.

5 Conclusions

In answer to seven key challenges associated with the development of AI for good, we propose a novel, systematic and responsible approach to the co-design and co-evaluation of intelligent recommender systems. Our method integrates ethical principles and software engineering best practice throughout system development, enhancing the finished product's potential for social good. We demonstrate the strengths of our approach on the iU&Mi urban mobility case study, where we combine, adjust and innovatively improve established development methodologies (SMART, PESTLE, MoSCoW) and participatory research best practice (GROW, TREQ) to create and conduct a series of engaging, collaborative activities over three workshops. These events gathered representatives of iU&Mi's target audience and the experts tasked with developing it who followed our principled process to co-design the recommender system's architecture and co-evaluate it, yielding (1) a system blueprint that creatively combines complementary components from across the AI tech spectrum and (2) a rigorous evaluation framework with calculated metrics assessing the blueprint's quality. Our contributions, i.e., the participatory process and the two workshop outputs, form the basis for the development of a fully articulated methodology that prescribes the responsible development of AI for social good, which the authors will complete in the next phase of their work.

Acknowledgments

This work was supported by Aston University's Pump Priming Fund (Project Code 20616-09). Special thanks to all workshop participants.

References

[1] David Adkins, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. 2022. Method Cards for Prescriptive Machine-Learning Transparency. In *2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN)*. 90–100. doi:10.1145/3522664.3528600

[2] Rachad Alao, Miranda Bogen, Jingang Miao, Ilya Mironov, and Jonathan Tannen. 2021. How Meta is working to assess fairness in relation to race in the US across its products and systems. *Meta Technical Report* (2021).

[3] Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R. Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T. Powers, Neil Urquhart, and Simon Wells. 2018. Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. *IEEE Technology and Society Magazine* 37, 4 (2018), 76–83. doi:10.1109/MTS.2018.2876107

[4] M S Muthu Selva Annamalai, I Bilogrevic, and E De Cristofaro. 2024. FP-Fed: Privacy-Preserving Federated Detection of Browser Fingerprinting.

[5] May Britt Bjerke and Ralph Renger. 2017. Being smart about writing SMART objectives. *Evaluation and Program Planning* 61 (2017), 125–127.

[6] D J Bogiatzis-Gibbons. 2024. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *Fairness, Accountability, and Transparency Conf.* ACM, New York, NY, USA, 74–84.

[7] B Braunschweig, S Buijsman, F Chamroukhi, F Heintz, F Khomh, J Mattioli, and M Poretschkin. 2024. AITA: AI trustworthiness assessment. *AI and Ethics* 4, 1 (2024), 1–3.

[8] Ingrid Burkett. 2012. An introduction to co-design. *Sydney: Knode* 12 (2012), 12.

[9] Dawn Chandler and Bill Torbert. 2003. Transforming inquiry and action: Interweaving 27 flavors of action research. *Action research* 1, 2 (2003), 133–152.

[10] Wesley Hanwen Deng, Solon Barocas, and Jennifer Wortman Vaughan. 2025. Supporting Industry Computing Researchers in Assessing, Articulating, and Addressing the Potential Negative Societal Impact of Their Work. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2025)*.

[11] Kate Donahue, Sreenivas Gollapudi, and Kostas Kollias. 2024. When are Two Lists Better than One?: Benefits and Harms in Joint Decision-Making. In *Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence (AAAI-24)*.

[12] Michael Gerlich. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15, 1 (2025). doi:10.3390/soc15010006

[13] Mélanie Gornet, Simon Delarue, Maria Boritchev, and Tiphaine Viard. 2024. Mapping AI ethics: a meso-scale analysis of its charters and manifestos. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 127–140.

[14] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4, 3 (2021), 324–332.

[15] M L Hilton and N J Cooke. 2015. Enhancing the effectiveness of team science. (2015).

[16] K D Key, D Furr-Holden, E Y Lewis, R Cunningham, M A Zimmerman, V Johnson-Lawrence, and S Selig. 2019. The continuum of community engagement in research: a roadmap for understanding and assessing progress. *Progress in community health partnerships: research, education, and action* 13, 4 (2019), 427–434.

[17] Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 100–112.

[18] D Kowald, S Scher, V Pammer-Schindler, P Müllner, K Waxnegger, L Demelius, A Fessl, M Toller, I G Mendoza Estrada, I Šimić, V Sabol, A Trügler, E Veas, R Kern, T Nad, and S Kopeinik. 2024. Establishing and evaluating trustworthy AI: overview and research challenges. *Frontiers in Big Data* 7 (2024).

[19] Seth Lazar and Alondra Nelson. 2023. AI safety on whose terms? *Science* 381, 6654 (2023), 138–138. doi:10.1126/science.adi8982

[20] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM.

[21] Peter R. Lewis and Stephen Marsh. 2022. What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research* 72 (2022), 33–49. doi:10.1016/j.cogsys.2021.11.001

[22] Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power hungry processing: Watts driving the cost of ai deployment?. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. 85–99.

[23] Ji-Ye Mao, Karel Vredenburg, Paul W Smith, and Tom Carey. 2005. The state of user-centered design practice. *Commun. ACM* 48, 3 (2005), 105–109.

[24] Elio Masciari, Areeba Umair, and Muhammad Habib Ullah. 2024. A systematic literature review on ai based recommendation systems and their ethical considerations. *IEEE Access* (2024).

[25] Alison Mathie and Gord Cunningham. 2003. From clients to citizens: Asset-based community development as a strategy for community-driven development. *Development in practice* 13, 5 (2003), 474–486.

[26] Eduardo Miranda. 2022. Moscow Rules: A Quantitative Exposé. In *Agile Processes in Software Engineering and Extreme Programming*, Viktoria Stray, Klaas-Jan Stol, Maria Paasivaara, and Philippe Kruchten (Eds.). Springer, Cham, 19–34.

[27] Nick Pangakis and Sam Wolken. 2025. Keeping Humans in the Loop: Human-Centered Automated Annotation with Generative AI. In *Web and Social Media*.

[28] Alina Patelli, John Rego Hamilton, Victoria Lush, and Aniko Ekart. 2022. A GENTLER Approach to Urban Traffic Modelling and Prediction. In *2022 IEEE Congress on Evolutionary Computation (CEC)*. 1–8.

[29] M.S. Reed. 2018. *The Research Impact Handbook*. Fast Track Impact.

[30] Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. 2024. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? arXiv:2407.21792 [cs.LG]

[31] Garima Sahu and Loveleen Gaur. 2024. Decoding the Recommender System: A Comprehensive Guide to Explainable AI in E-commerce. In *Role of Explainable Artificial Intelligence in E-Commerce*. Springer, 33–52.

[32] Rahul Singh. 2024. The Recommender System Paradox: Autonomy in the Age of AI Recommendations. In *2024 3rd International Joint Conference on Information and Communication Engineering (JCICE)*. IEEE, 139–143.

[33] techUK. 2025. *Mapping the Responsible AI Profession: Current Practice and Future Pathways*. Technical Report.

[34] Jan Walmsley. 2006. Normalisation, emancipatory research and inclusive research in learning disability. In *Overcoming disabling barriers*. Routledge, 335–355.

[35] Shoujin Wang, Xiuzhen Zhang, Yan Wang, and Francesco Ricci. 2024. Trustworthy recommender systems. *ACM Trans on Intel. Sys. and Tech.* 15, 4 (2024), 1–20.

[36] YouTube [n. d.]. https://www.youtube.com/watch?v=l8Va6A7mKRQ&t=1s

[37] E Zhang, R Khurana, and V Khare. 2025. Governance, Risk and Compliance (GRC) Engineering: Data, AI, Automation, and the Future of Compliance to Audits.

[38] R V Zicari, J Brodersen, J Brusseau, B Düdder, T Eichhorn, T Ivanov, G Kararigas, P Kringen, M McCullough, F Möslein, et al. 2021. Z-Inspection®: a process to assess trustworthy AI. *IEEE Trans. on Technology and Society* 2, 2 (2021), 83–97.

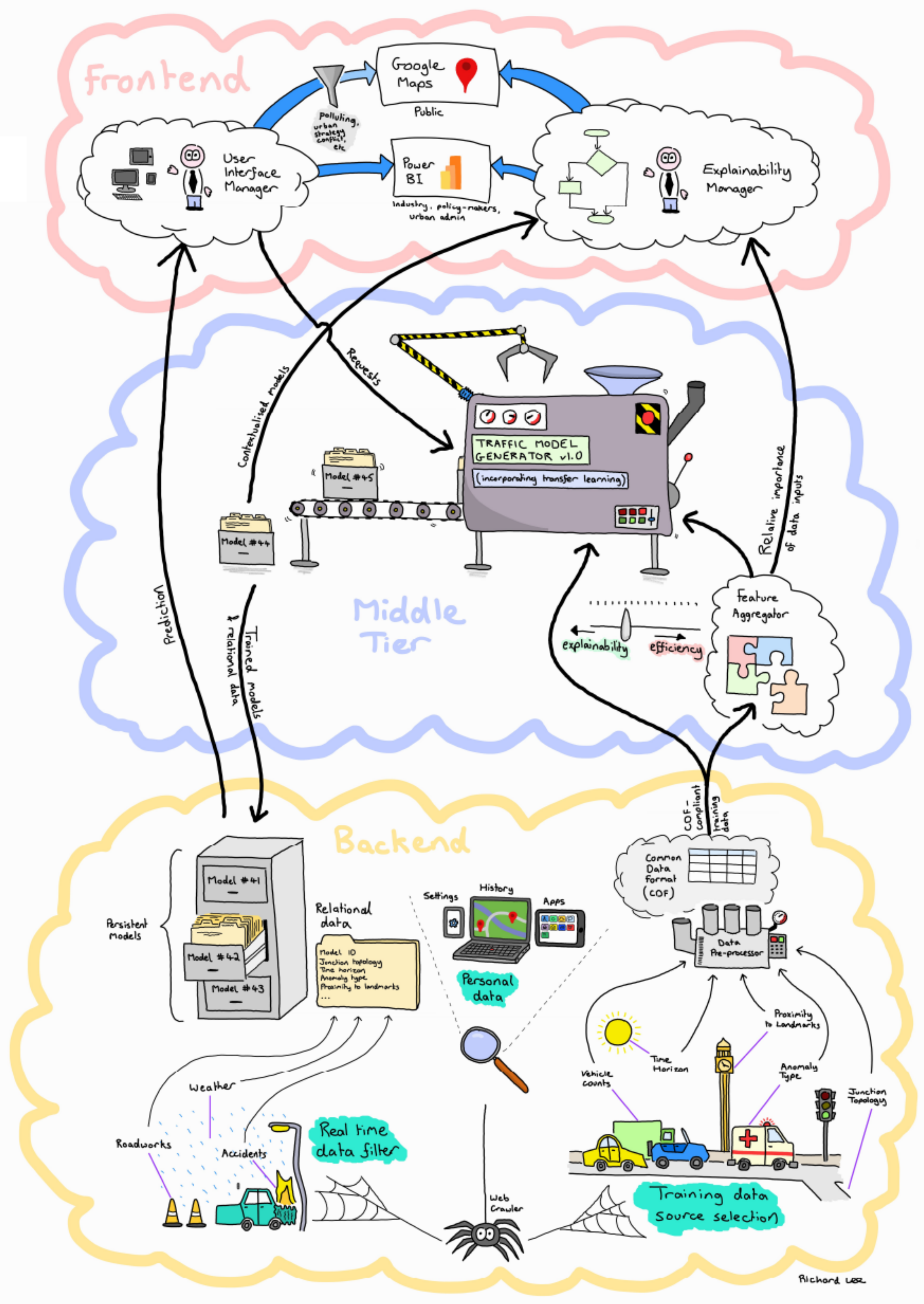


Figure 5: The iU&Mi architecture diagram in a non-technical rendition. Logically equivalent to the UML diagram in Fig. 4, it provides the iU&Mi audience with an intuitive view of the recommender system's inner-workings.