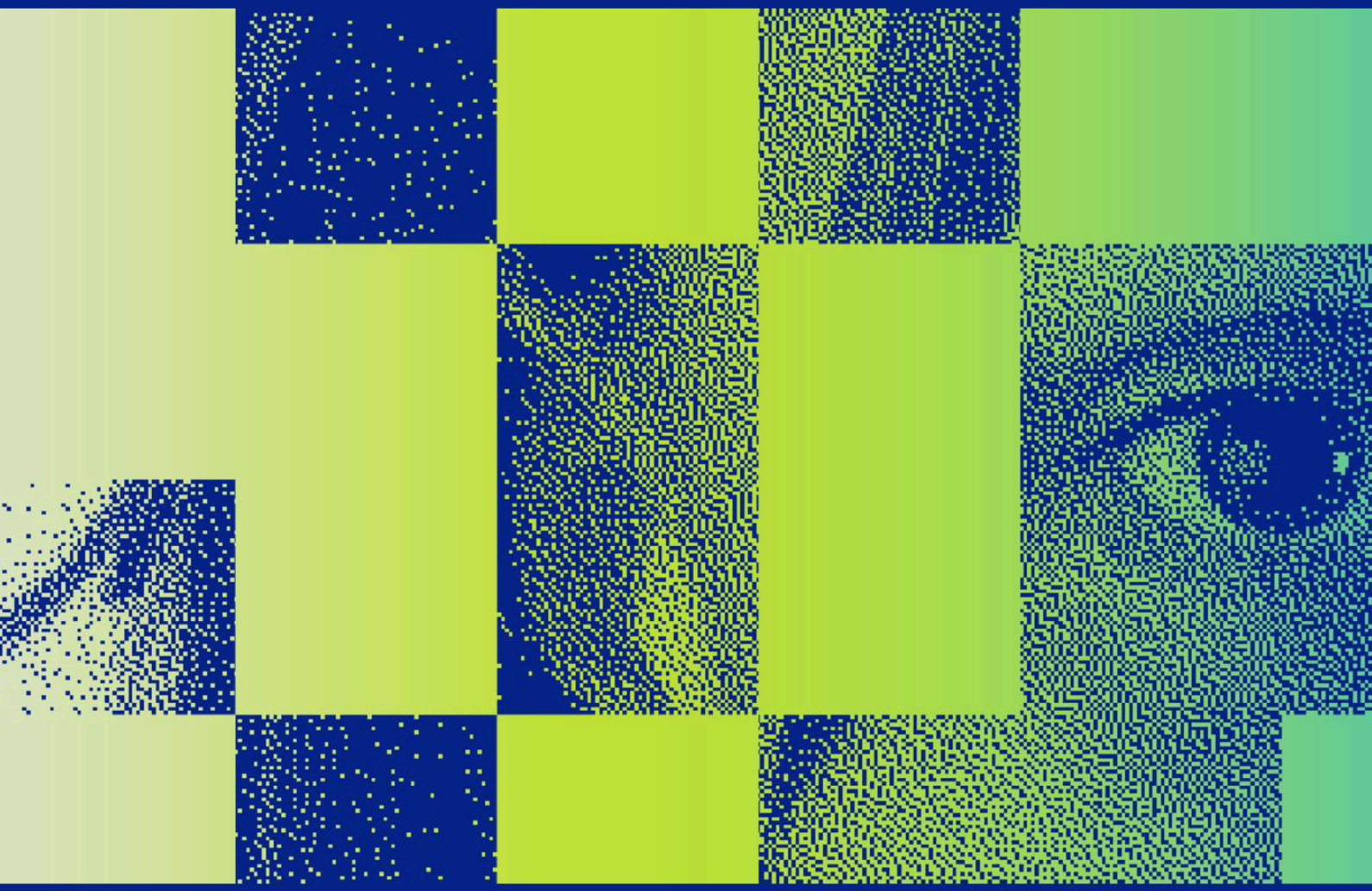

Building a user-centric AI data transparency approach

ODI Research
ADVANCING TRUST IN DATA

December 2024

Contents

About 2
Executive summary **3**
Findings 4
Recommendations 4
Introduction **6**
Contribution 8
Designing a user-centric AI Data Transparency Index **9**
a) User needs for AI data transparency 9
b) Establishing the dimensions needed by these users 11
c) Using the Index 14
Findings **17**
Assessing data transparency of 22 models 17
Discussion **22**
Assessment method 22
Recommendations 23
Future work 24
Limitations of approach 26
Annex A: User needs table **27**
Annex B: Models assessed **28**

About

This report was researched and produced by the Open Data Institute (ODI) and published in December 2024. Its lead authors were Ben Snaith, Sophia Worth and Elena Simperl, with additional support from Neil Majithia.

We are grateful to the Department of Informatics at King's College London for providing support with the assessments. We would also like to thank participants of a workshop we held as part of the ODI's Data Centric AI (DCAI) event in October 2024.

If you want to share feedback by email or would like to get in touch, contact the DCAI programme lead Elena Simperl at elena.simperl@theodi.org.

Executive summary

Despite broad agreement on the critical importance of data to AI development, attempts to understand and query this data are stilted by poor transparency practices. Recognising the fears of a 'growing data transparency crisis' for AI[1] necessitates an investigation into how transparency practices do not meet the diverse needs of the responsible AI ecosystem. As developers frequently fail to disclose details of their training datasets[2] transparency practices vary significantly, ranging from widespread opacity to clear and detailed disclosures. This inconsistency hampers efforts to ensure fairness, identify biases, and comply with regulations, leaving researchers, policymakers and the public unable to make informed decisions about AI systems or conduct large-scale comparative research.

To provide a systematic exploration of how aligned data transparency practices are with user needs, the Open Data Institute (ODI) has developed the AI Data Transparency Index (AIDTI), a maturity assessment framework designed to evaluate the level of data transparency across AI models. This approach considers transparency of the data processes that take place in upstream AI development. Grounded in the needs of two primary users – developers and Responsible AI (RAI) researchers – the Index assesses transparency across several upstream dimensions:

1. **List of datasets used:** Information about the origin and composition of datasets.
2. **Data collection method:** Clarity on how data was gathered and any potential biases introduced.
3. **Pre-processing information:** Documentation of steps taken to prepare data for training.
4. **Accessible transparency information:** Whether standardised and open documentation approaches are used.
5. **Copyrighted and personal data in data:** Disclosure of whether sensitive material has been included in training.
6. **Environmental impact:** Insights into energy consumption.
7. **Human and organisational supply chains:** Transparency about the labour and entities involved in data preparation.

[1] Longpre, S. et al (2024) 'Consent in Crisis: The Rapid Decline of the AI Data Commons'.

[2] Schaul, K, Chen, SY, Tiku, N (2023) 'Inside the secret list of websites that make AI like ChatGPT sound smart'.

By adopting a user-centric approach, the AIDTI goes beyond assessing the presence of information, emphasising its quality, accessibility and relevance to specific user groups. This report presents findings from the first version of the AIDTI, with the assessment of 22 AI models. Models were evaluated using the AIDTI framework, scoring their maturity on a scale from low to high:

The findings of this piece highlight the distinct lack of purpose underlying how transparency information is shared (if at all). It does this through establishing the range of user needs, and then assesses how 22 models fulfill the needs of two of the primary users of transparency information.

Findings

Overall:

- **High maturity:** Demonstrated by five model providers, characterised by detailed, accessible documentation, consistent use of transparency tools, and a proactive approach to explaining decisions made in the development process.
- **Medium maturity**: Six model providers met some transparency criteria but lacked consistency for all dimensions.
- **Low maturity**: Eleven model providers shared limited or poor-quality information, suggesting a general reluctance to be open.

Alongside this:

- Aspects of the data lifecycle, such as data sources, collection methods and pre-processing activities, were more consistently documented.
- The environmental impact of models is beginning to be meaningfully documented as model providers reckon with the energy usage of AI.
- Information on the human supply chain, and the inclusion of copyright or personal data in AI models, was consistently poorly documented.

Recommendations

Recommendations for key stakeholders:

- Creating a comprehensive ‘gold standard’ for developers to provide transparency information for diverse stakeholder needs.
- Consistently drive minimum transparency standards, with a focus on creating balanced documentation requirements that do not overly burden model providers.
- Continue research and advocacy efforts to hold AI developers accountable by pushing for higher quality, and more meaningful, transparency information.

AI Data Transparency Index recommendations:

- Expand the AI Data Transparency Index by assessing more models across different jurisdictions, sectors, and use cases.
- Assess whether transparency is aligned with the needs of other key users of transparency information.
- Create an interactive, machine-readable system for consistently documenting and sharing AI transparency information.

The AIDTI represents a significant step towards more meaningful and user-centric AI data transparency across the ecosystem, not just by evaluating practices as they are, but through providing a framework for the direction of travel to a more meaningful data transparency approach. Future work will focus on expanding the AIDTI to include more models and stakeholder perspectives, integrating machine-readable data for greater accessibility, and exploring interactive systems to visualise AI supply chains. In addition, this work aims to establish a new direction of research to ensure meaningful transparency to help all those ensuring that AI ecosystems are responsible and trustworthy.

Introduction

Data is a cornerstone of AI development. AI is often trained and fine-tuned using billions of datapoints scraped from the web, purchased in bulk, or contributed to by a vast number of human annotators. Knowing what is in the datasets used to train models, and how they have been compiled, is vitally important for the development and deployment of safe and responsible AI systems. AI data transparency refers to the openness about how data is utilised throughout the AI lifecycle[3], with a focus on upstream data components: training data, fine-tuning, reference data and benchmarks.[4] Despite the importance of data, most leading AI firms have been unwilling to disclose details about the datasets used to train and test their models[5], contributing to what has been termed a 'growing data transparency crisis'.[6] The Stanford Foundation Model Transparency Index, which assesses the major foundational models that provide the backbone of many AI tools and services, demonstrated that transparency regarding the data used was very low compared to other aspects of transparency[7]. Recent ODI research examined data transparency across a range of models linked to recent 'AI incidents' highlighted in the media and identified a similarly low presence of data transparency information, alongside key barriers for accessing this information.[8]

Initiatives are being undertaken across multiple sectors, stakeholders and contexts to attempt to tackle this issue. These include draft regulations emerging across a variety of jurisdictions, voluntary agreements such as the US' Executive Order in 2023[9], technical standards to help developers better address transparency needs, like the Croissant ML data standard[10], and the uptake of AI documentation approaches for AI transparency at the level of datasets (like datasheets[11], nutrition labels[12], data statements[13]),

[3] Hardinges, J and Simperl, E. (2024), 'A data for AI taxonomy'.
[4] ibid, the 'Developing AI systems' section.
[5] Schaul, K, Chen, SY, Tiku, N. (2023), 'Inside the secret list of websites that make AI like ChatGPT sound smart'.
[6] Longpre, S. et al (2024), 'Consent in Crisis: The Rapid Decline of the AI Data Commons'.
[7] Bommasani, R. et al. (2024), 'The Foundation Model Transparency Index'.
[8] Worth, S. et al (2024), 'AI data transparency: an exploration through the lens of AI incidents'.
[9] The White House (2023), 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'.
[10] Majithia, N., Carey-Wilson, T., Simperl, E. (2024), 'Transforming AI data governance with Croissant: a new standard for ML metadata'.
[11] Gebru, T. et al. (2018), 'Datasheets for Datasets'.
[12] Holland, S. et al. (2018), 'The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards'.
[13] Bender, E.M. and Friedman, B. (2018), 'Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science'.

models[14], and even at AI 'system' level[15]. These approaches guide those involved in AI development to share transparency information publicly. Still, recent research has demonstrated that AI documentation approaches are often used inconsistently, if at all, by those delivering AI systems[16].

As adoption of all these approaches steadily grows, it is important to keep sight of the fact that transparency is not the end objective. Instead, transparency is required to improve responsible decision-making at all stages of the AI lifecycle – for example, in helping everyone from developers to deployers to affected individuals to prevent racially biased AI systems from entering and causing harm in the public domain. It is clear that without transparency information, a variety of needs will not be met. A few examples are:

- Ability of developers, researchers and ethicists to understand and address biases or remove harmful content from training data.
- Lawmakers' ability to understand whether foundation models have ingested personal data or copyrighted material.
- Users' and deployers' ability to trust or contest systems they are relying on if they know how they have been developed.

It remains unclear how far away we currently are from meeting the needs of users who need transparency information for a whole variety of purposes – from development to compliance, from ethical decision-making to introducing new policy.

To understand what a more user-centric approach to data transparency should resemble, we are working to design an AI Data Transparency Index (AIDTI) to understand the current level of transparency adoption, with three main questions at its core:

- Who is data transparency for?
- What are the main use cases for data transparency?
- What are the barriers to accessing data transparency?

In this report, we set out an initial set of user personas and use cases based on exploratory research, and share our Index methodology. We drew insights about

[14] Liang, W., Rajani, N., Yang, X. et al. (2024), 'Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI'.
[15] Arnold, S., Yesilbas, D., Gröbner, R., Riedelbauch, D., Horn, M., Weinzierl, S. (2024), 'Documentation Practices of Artificial Intelligence'.
[16]Liang, W. et al. (2024), 'Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI'.

the current landscape of transparency related to the established needs of ‘responsible AI researchers’ and developers, identifying inconsistent maturity in how developers share necessary transparency information, with a stark difference between developers who are evidently committed to openness and transparency and those favouring extremely opaque practices. However, in the future, we look to build further on our methodology and understand the transparency landscape for a wider range of user groups.

Contribution

Through this, we have a number of contributions to the burgeoning AI data transparency space:

- Personas detailing the needs and requirements of AI data transparency.
- Use cases of how this transparency information could be used by these users.
- A maturity assessment methodology to create the Index.
- Research findings, insights and recommendations from evaluating 22 AI models.

Designing a user-centric AI Data Transparency Index

In recognition of the need for more meaningful data transparency across AI lifecycles, we designed a user-focused research project. This work builds upon interventions that the ODI has already made into this space – establishing the need for more data transparency in AI[17], and evaluating transparency through the lens of reported incidents.[18] We continued our research into literature on AI, model and supply chain transparency. To develop the user personas, we interviewed three experts that matched the different personas, and undertook a user needs workshop to build our understanding of the other personas' needs. When we talk about user needs, we are considering how individuals who match these personas would likely look for, interact with, and use transparency information. The AIDTI uses a maturity assessment to consider not only what information about AI data is being shared, but also to establish a broader direction of travel for the whole field of AI data transparency.

a) User needs for AI data transparency

The design of our AIDTI is grounded in our own observations and feedback from researchers and practitioners in the wider field of responsible AI. As noted earlier, while there is a growing body of evidence for the importance of AI data transparency, and some technical and socio-technical solutions, AI providers are mostly still distant from delivering best practice. One of the reasons for this is the gap between what the solutions deliver, and the needs of their potential users; perhaps unsurprisingly for an emerging topic, so far most efforts have focused on the presence of information rather than its practical utility for different user groups. For example, prior research has looked at the quality of information shared via data cards on Hugging Face.[19]

---

[17] Snaith, B. et al (2024), 'Policy intervention 1: Increase transparency around the data used to train AI models'.
[18] Worth, S. et al (2024), 'AI data transparency: an exploration through the lens of AI incidents'.
[19] Yang, X., Liang, W. and Zou, J. (2024), 'Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on HuggingFace'; Liao, Q.V., Subramonyam, H., Wang, J. and Wortman Vaughan, J. (2023), 'Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience'.

The demand for transparency to be tailored towards contextual user needs is detailed through empirical research[20], which establishes the different decision-making processes that individuals take across different contexts – such as in policymaking, responsible research academia or technical development. We engaged with three people via interviews and 45 industry experts through a workshop to identify data transparency personas.

This work established seven key personas for data transparency information:

- **Data scientists and developers** who require data transparency to reproduce and fork models
- **RAI researchers** who are interested in answering issues of fairness, bias and data practices and therefore require access to information about training data, data augmentation and similar
- **Policymakers and regulators** who are responsible for ensuring safe and responsible AI use, and need to understand existing practices to ensure that regulatory obligations (such as the EU AI Act) are being followed, or propose new regulation or agreements to improve practices
- **Members of the public** who wish to understand which AI models and tools are safe and ethical
- **Creatives** who wish to understand if their copyright or intellectual property data has been used in training certain models[21]
- **Journalists** who want to expose potential harms in the training and use of AI, and therefore need to be able to follow the data across the whole lifecycle
- **Lawyers** who may be defending or prosecuting alleged copyright infringement or inappropriate data processing within training models.

The full personas, including the use cases for the data, example applications and information needs are included in Annex A.

---

[20]Norval, C. et al. (2022), 'Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability'; Schor, B.G.S. et al. (2024), 'Mind The Gap: Designers and Standards on Algorithmic System Transparency for Users'.

[21] See: Hardinges, J. et al (2024), 'Policy intervention 2: Update our intellectual property regime to ensure AI models are trained fairly'.

b) Establishing the dimensions needed by these users

Building on these personas, we explored specific user needs and sought to establish the core criteria ('dimensions') of meaningful data transparency. These dimensions were designed to meet the requirements of our primary use case for this assessment: developers and RAI researchers. Although their needs differ, these two user groups both required a detailed understanding of 'upstream' elements of AI development, and therefore appear to have similar transparency needs.. The presence of previous research on the user needs of these groups allows us to test our approach and its appropriateness, and ensure that our findings are comparable to previous research.

Dimensions:	Explanation for developers and RAI researchers
1) List of datasets	Provides the ability to trace the data throughout the AI lifecycle and assess governance controls[22] and to understand pre-trained models and how to develop these responsibly.[23] This list of datasets can be useful to users who wish to replicate the models, or understand concerns over data quality.[24]
2) How was the data collected?	Allows others to repeat the collection and assess the reliability of the model or repeat for their own developments. Researchers need to understand whether the data collection methodology introduced any bias issues that affect the model's reliability or fairness, either through provided transparency information[25] or data and algorithmic auditing.[26] [27] It can also help to understand any challenges with data collection (such as whether copyrighted or personal data has been used).

[22] Carey-Wilson, T. et al (2024), 'Understanding data governance in AI: A lifecycle perspective'.
[23] Liao, Q.V., Subramonyam, H., Wang, J. and Wortman Vaughan, J. (2023), 'Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience'.
[24] Sambasivan, N. et al (2021), 'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI'.
[25] Tawakuli, A. and Engel, T. (2024), 'Make Your Data Fair: A Survey of Data Preprocessing Techniques that Address Biases in Data Towards Fair AI'; ICO, n.d 'How do we ensure fairness in AI?'.
[26] Yanisky-Ravid, S. and Hallisey, S. (2018), 'Equality and Privacy by Design': Ensuring Artificial Intelligence (AI) Is Properly Trained & Fed: A New Model of AI Data Transparency & Certification As Safe Harbor Procedures'.
[27] Eticas, n.d., 'Guide to AI Auditing'.

3) What pre-processing activities took place?	Allows auditing of the AI model and an assessment of whether the model has suitable frameworks to account for issues with the dataset regarding accuracy, quality, and harmful content.[28] For both researchers and developers, an understanding of pre-processing supports efforts for safety and reproducibility.[29]
4) Are accessible mechanisms for transparency used?	Transparency regarding data practices (and identification of limitations within the data) allows others to make informed decisions about whether to use or repeat the model. Using transparency mechanisms can enable a consistency of approach across models and developers and ensure that large-scale research studies can take place.[30]
5) Was copyrighted data used in the AI model? 6) Was personal information used in the AI model?	There are specific legislative requirements related to documenting whether/how personal data was used.[31] Major concerns and legal challenges[32] have been raised over the use of copyrighted data within AI systems, particularly generative AI systems, which has lead to lobbying for legal requirements on companies to disclose the copyright status of their data. Developers need to be able to ensure that they know if this information has been included in training[33], ensuring that they are processing data in compliant manner across jurisdictions,[34] and researchers need to be able to validate these claims.[35]
7) Environmental impacts of model	Understanding the training resources required provides insight into the environmental and financial costs of the AI system, allowing developers to make decisions regarding compute and energy costs if they are looking to replicate the model, particularly as access to compute is becoming more challenging.[36] [37] Understanding environmental impacts supports responsible AI assessments, as the energy costs of using large-scale datasets for training are evaluated.[38]

[28]Liao, Q.V., Subramonyam, H., Wang, J. and Wortman Vaughan, J. (2023), 'Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience'.
[29] Tawakuli, A. and Engel, T. (2024), 'Make your data fair: A survey of data preprocessing techniques that address biases in data towards fair AI'; Saplicki, C, Bante, M (2023), 'Fairness in Machine Learning: Pre-Processing Algorithms'.
[30] Liang, W. et al. (2024), 'Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI'
[31] ICO (2023), 'Joint statement on data scraping and the protection of privacy'.
[32] Vincent, J (2022), 'The scary truth about AI copyright is nobody knows what will happen next'.
[33] ICO (2024), 'ICO urges all app developers to prioritise privacy'.
[34] Lu, Q. et al. (2023), 'Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering'.
[35] Leffer, L. (2023), 'Your Personal Information Is Probably Being Used to Train Generative AI Models'.
[36] Shearer, E., Davies, M., Lawrence, M. (2024), 'The role of public compute'.
[37] Kudiabor, H. (2024), 'AI's computing gap: academics lack access to powerful chips needed for research'.
[38] Wu, C.-J. et al. (2021), 'Sustainable AI: Environmental Implications, Challenges and Opportunities'.

8)	Human and organisational supply chain	AI lifecycles and supply chains are typically multi-stakeholder. For accountability, and to investigate fairness across the supply chain, it is necessary to be able to identify the people and organisations involved in activities such as data collection, filtration, augmentation and fine-tuning.[39] This transparency can be useful in identifying fairness and ensuring labour rights are respected.[40] But given the lack of transparency around the data used to train many of the popular AI models[41], organisations may not even be aware of how reliant on this labour they are.[42] For developers, this can be useful to explore potentials for collaboration and best practice across the supply chain, and identify opportunities for efficiency.[43]

Table 1: User needs table and the information required

We adopted a maturity assessment – rather than a yes/no assessment – to support the needs of specific user groups. Alongside assessing whether developers have published certain information about the data they used, our assessment also considered whether methodologies were detailed, and whether there was context and explanation for the decisions made throughout the AI lifecycle. As we wrote in a preliminary analysis of AI data transparency, there is a need to see 'how well those developing, deploying and using AI systems understand biases, limitations and legal obligations associated with use of this data, to ensure systems are implemented appropriately'.[44] Maturity assessment allows us to capture that contextual, descriptive information, in addition to documenting whether the information was available. This design decision is also in line with broader attempts elsewhere in the AI community to bring greater explainability to foster trust.[45]

For the same reasons, this style of assessment also helps to future-proof assessment results against potential open-washing[46]: our approach does not just require that information be published to achieve good results, but that it is published in a form truly appropriate for its users – and therefore is not as easily gamed for a higher assessment result. Further, a maturity approach appreciates that transparency is not simply 'a precise end state in which everything is clear and apparent' but 'a system of observing and knowing that promises a form of control'.[47]

[39] Muldoon, J., Graham, M., & Cant, C. (2024.). 'Feeding the Machine: The Hidden Human Labour Powering AI'.
[40] Ustek Spilda, F., Brittain, L., Cant, C., & Graham, M. (2024), 'The Unmagical World of AI: Workers at the bottom of the AI supply chain'.
[41] Snaith, B. et al (2024), 'Policy intervention 1: Increase transparency around the data used to train AI models'.
[42] Hardinges, J. et al. (2024), 'Policy intervention 3: Enforcing people's rights in the data supply chain'.
[43] Suryadevara, M., Rangineni, S., Venkata, S. (2023), 'Optimizing Efficiency and Performance: Investigating Data Pipelines for Artificial Intelligence Model Development and Practical Applications'.
[44]Worth, S. et al (2024), 'AI data transparency: an exploration through the lens of AI incidents'.
[45] People + AI Research team (2021), 'Explainability + Trust, People + AI Guidebook'.
[46] Liesenfeld, A. and Dingemanse, M. (2024), 'Rethinking open source generative AI: open-washing and the EU AI Act'.
[47] Ananny, M. and Crawford, K. (2018), 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability'.

For some of the dimensions, technical knowledge is required to interpret the language and intentions in release notes. Historically, the data used to build AI systems has commonly been viewed as a 'technical' element, and therefore the details of data used have been relevant only to those with technical expertise.[48] As release notes are designed for a technical audience, the explanations and insights might be ineffective or counterproductive towards trust with other audiences.[49] With this AIDTI, we make a first step towards changing this, so that transparency information is accessible and useful to a range of different technical and non-technical audiences, to ensure that it can achieve its purpose.

c) Using the Index

To assess whether providers of foundational models were meeting the needs of our first two user groups – developers and RAI researchers – we tested our methodology on 22 models.

[48] Jarrahi, M.H., Memariani, A., Guha, S. (2023), 'The Principles of Data-Centric AI'.
[49] Zieglmeier, V. and Pretschner, A. (2021), 'Trustworthy Transparency by Design'.

Model developer	
Model name	
Assessor	
Assessor role	

Data type	Explanation	Maturity Assessment Score	Assessor notes/justification	requested information	Links to transparency information/sources
List of datasets	Ability to trace the 'data supply chain' - many of the other questions depend on the availability of more granular data. This includes details of data used during training, validation, fine-tuning, benchmarks	3			
How was the data collected?	Allows others to repeat the collection and assess the reliability of the model or repeat for their own developments. There is a need to understand whether the data collection methodology introduced any bias issues that affect the model's reliability or fairness. It can also help to understand any challenges with data collection (such as whether copyrighted or personal data has been used).	3			
What pre-processing activities took place?	Allows auditing of the AI model and an assessment on whether the model has suitable frameworks to account for issues with the dataset re: accuracy, quality, harmful content.	2			
Are accessible mechanisms for transparency used?	Transparency regarding data practices (and identification of limitations within the data) allows others to make informed decisions about whether to use or repeat the model.	1			
Was copyrighted data used in the AI model?	There are specific legislative requirements related to documenting whether/how copyrighted data was used.	1			
Was personal information used in the AI model?	There are specific legislative requirements related to documenting whether/how personal data was used.	2			
Environmental impacts of model	Understanding the training resources required provides insight into the environmental and financial costs of the AI system, allowing others to query its efficiency, understand costs and assess possibility of replication.	2			
Human and organisational supply chain	AI lifecycles and supply chains are typically multi-stakeholder. For accountability and to investigate fairness across the supply chain, it is necessary to be able to identify the people and organisations involved in activities such as data collection, filtration, augmentation and fine-tuning. This transparency can be useful in identifying fairness and ensuring labour rights are respected.	1			
	Score	15			

Figure 1: Example blank assessment scorecard. Full maturity criteria were provided to the assessor.

To select the models to assess, our starting point was companies that agreed to the Frontier AI Safety Commitments at the AI Seoul Summit 2024 to undertake the responsible development of AI.[50] We supplemented those with models that had been deemed to meet, or were close to meeting, the Open Source Initiative's Open Source AI definition.[51] This gave us a global cross-section of companies with a commitment to safety and openness, which one would assume would be predicated on strong transparency approaches. This also ensured that the models we assessed were from a wider cross-section to related work[52], including a more international perspective, with models from the USA, UAE, China and South Korea.

To carry out the assessment, we had the support of King's College London's Department of Informatics. We recruited 12 volunteers from the two user groups, data scientists and RAI researchers. We allocated one or two models per assessor. Each assessment took around an hour to complete. The models were scored based on our maturity methodology: low maturity meant they provided no or poor information for each metric and high indicated that they provided the information and explained their approach and methodology. The assessors were asked to look for officially authored technical papers, release notes and model and data cards to source the information. Based on the maturity given for each dimension (1 being low, 3 high), an overall score was given: 8-12 low, 13-17 medium, 18+ high. Each assessor compiled notes for their qualitative assessments to justify the maturity score given, including links to the material they sourced to carry out the assessment. These notes were analysed for trends and patterns, which we have explored further in the findings section.

Each result was validated by an ODI researcher, who checked that the assessors sufficiently justified the scoring given in accordance with the maturity criteria and correctly applied the methodology. Four models were also assessed blind for a second time. Following this validation process, a number of the scores were adjusted slightly to ensure consistency across the model provider assessments. One assessment strayed from the methodology and therefore the assessment was repeated in full by another assessor and re-validated.

[50] DSIT (2024), 'Frontier AI Safety Commitments, AI Seoul Summit 2024'.
[51] Open Source Initiative (n.d.), 'The Open Source AI Definition – 1.0'.
[52] Bommasani, R. et al. (2024), 'The Foundation Model Transparency Index'.

Findings

Assessing data transparency of 22 models

Overall, there was a significant spread of maturity across the 22 model providers:

- Five were ranked as high maturity for AI data transparency.
- Six were ranked as medium.
- Eleven were ranked as low maturity.

Several consistent best practices emerged from models with strong overall maturity for data transparency:

- Publication of the training dataset, or a detailed list of sources and citations for the datasets.
- A tone and style of report that makes a significant attempt to explain.
- A consistent use of model cards on Hugging Face, meaning that much of the requested information was produced in a quasi-standardised format.
- A technical report of the model and dataset, available open access through arXiv.

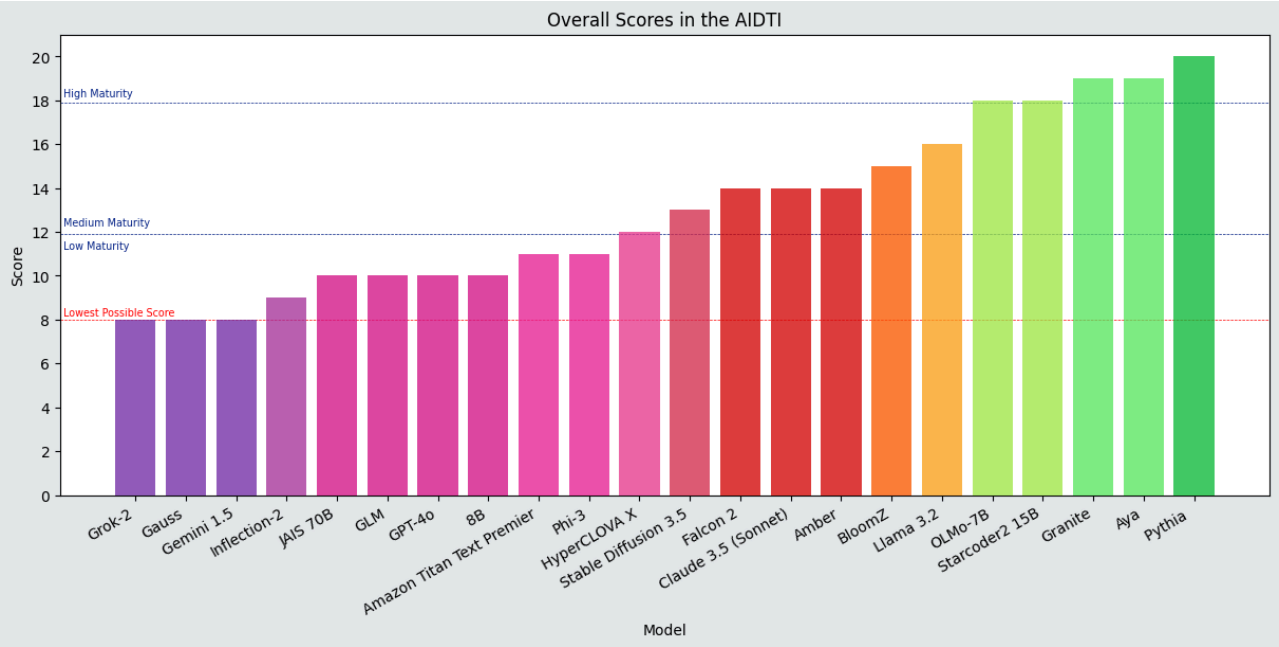

Figure 2: Overall scores from the AIDTIx assessment

The central question of our maturity assessment was whether the assessor is capable of finding and understanding the transparency information – this may mean that the information we have said is inaccessible in this assessment could be out there in some form. There is a tendency to 'release by blogpost'; where developers '[reap] the benefits of mimicking scientific communication… without actually doing the work'.[53] We deliberately did not validate the results with the model providers, as our aim was to explore to what extent this information is available and can be found and understood within reasonable effort by a third party. As all results were checked by a second researcher, any misalignment with what the model providers believe they share is likely down to poor findability or poor explanations, and should be improved for meaningful transparency.

Across all findings, we note that on average there was greater maturity in the top four categories than the bottom four (see Figure 3 below which shows the variance across the metrics). For example, publishing a list of datasets, and the pre-processing information, has been part of model transparency attempts since concerns over black-box algorithms ramped up in the 2010s.[54] Reflected in this is the connection between the developers that used a form of accessible transparency mechanism (mostly via a model card on Hugging Face) and the likelihood that they were also publishing information about data sources, collection methods and pre-processing activities. What is also interesting is the cluster of model providers that have used a model card, but were still not achieving higher maturity for transparency of data sources and collection methods – corroborating research on inconsistency in the amount of detail included within model and data cards.[55]

[53] Liesenfeld, A. and Dingemanse, M. (2024), 'Rethinking open source generative AI: open-washing and the EU AI Act'.
[54] Guidotti, R. et al. (2018), 'A Survey of Methods for Explaining Black Box Models'.
[55] Liang, W. et al. (2024), 'Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI'.

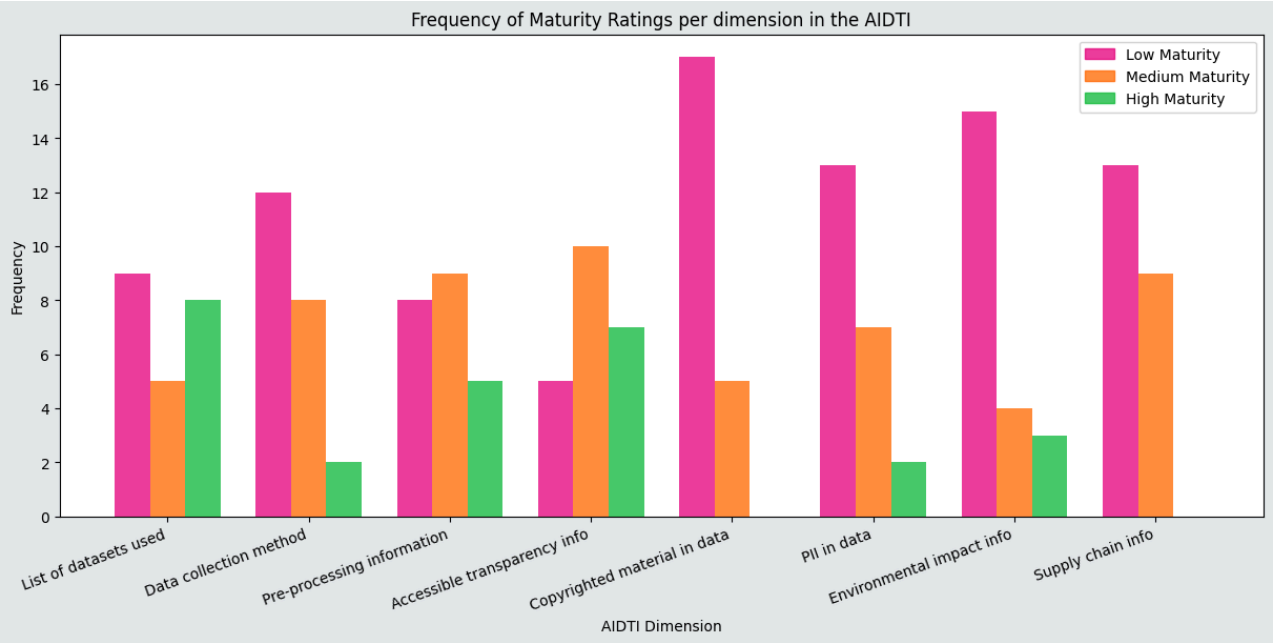

Figure 3: a comparison of maturity scores for each dimension of the AIDTI

We identified a similarly mixed-maturity level in how pre-processing activities were documented. As the Data Provenance Explorer establishes, most AI development happens through fine-tuning[56] and few-shot learning[57] of pre-trained models. For the models that were documented more thoroughly for this criteria, model providers included the detailed steps that were taken to make the dataset suitable for use. For example, Cohere documented how inappropriate language was filtered through a comparison with a list of obscene words on GitHub.[58] Similarly, Meta included thorough explanations of pre-processing activities in the documentation of Llama 3.2. [59] This included details of how harmful content was filtered, the de-duplication that took place, and how heuristic filtering was used to screen low-quality and repetitive data. For high maturity, we would expect clear documentation to demonstrate the many types of data needed to build, use and monitor AI systems safely and effectively.[60]

There was less consistency regarding whether either copyrighted information or personal information was used to train or fine-tune a mode. For both dimensions, these factors tended to be discussed briefly as part of discussions regarding data collection or pre-processing, but rarely warranted separate clear discussions. With significant attention on model providers to be clear about training data – via regulatory approaches and pressure from users and

[56] Bergmann, D. (2024), 'What is Fine-Tuning?'.
[57] IBM (n.d.), 'What Is Few-Shot Learning?'.
[58] Raffel, C. et al. (2019), 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer'.
[59] Grattafiori, A. et al. (2024), 'The Llama 3 Herd of Models'.
[60] Hardinges, J. and Simperl, E. (2024), 'A data for AI taxonomy'.

creatives – this will likely need to change; if we follow that data transparency is 'the ability of subjects to effectively gain access to all information related to data used in processes and decisions that affect the subjects'[61], then being able to identify the likelihood that personal or copyrighted information was used in the training of the model is foundational.[62]

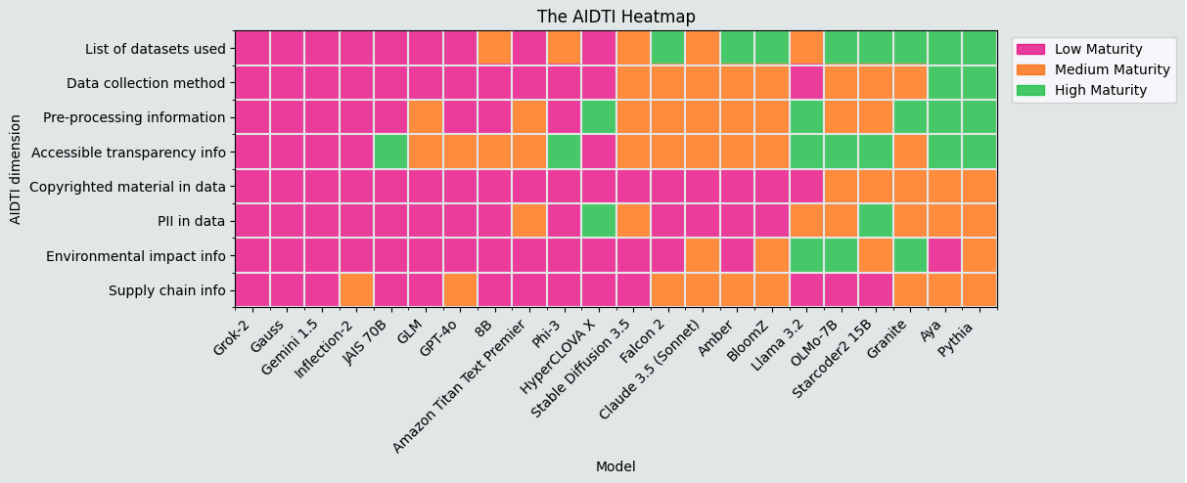

Figure 4: AIDTI heatmap demonstrating widespread inconsistencies in transparency documentation across 22 models

Connected to the discussion above, regarding disclosing the use of copyrighted data[63], we also expect recognition of the organisations and labour involved across the AI supply chain to increase.[64] None of the assessed models demonstrated high maturity in transparency regarding this aspect. With a steady stream of allegations emerging from the sector of labour abuses[65] and exploitation of creatives[66], the ability for RAI researchers especially to be able to follow the data throughout the AI lifecycle and evaluate for safety and responsibility is vital. While 'not all AI relies on crowdwork, and not all crowdwork feeds into AI,' the two are 'inextricably linked'[67]. There is significant human involvement behind the data used to train foundation AI models, such as collecting data, filtering and moderating it, and labelling it.[68] Given the issues with fairness in the AI supply chain, transparency regarding the role of data workers is a necessity, even if, for

[61] Barhamgi, M. and Bertino, E. (2022), 'Editorial: Special Issue on Data Transparency—Uses Cases and Applications'.
[62] Hardinges, J. et al. (2024), 'Policy intervention 2: Update our intellectual property regime to ensure AI models are trained fairly'; Hardinges, J. et al. (2024), 'Policy intervention 3: Enforcing people's rights in the data supply chain'.
[63] Hardinges, J. et al. (2024), 'Policy intervention 2: Update our intellectual property regime to ensure AI models are trained fairly'.
[64]Hardinges, J. et al. (2024), 'Policy intervention 3: Enforcing people's rights in the data supply chain'.
[65] Perrigo, B. (2023), 'OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic'.
[66] Boran, M. (2024), 'OpenAI's Sora Leaked Online Over 'Unpaid Labor,' Artists Say'.
[67] Gonzalez-Cabello, M. et al. (2024), 'Fairness in crowdwork: Making the human AI supply chain more humane'.
[68] Estampa (2024), 'Cartography of generative AI'.

example, AI itself is used for text annotation tasks.[69] We expect this data to also be of high interest to other users we identified - the public, journalists and policymakers – as concern over these practices grows.

The environmental impact of data and AI is receiving greater attention, with nearly half of the model providers we analysed (nine in 22) including some information regarding energy usage, emissions or environmental impact of the model. This is an encouraging trend. However, a recent study of model documentation practices has found that only 2% of model cards on Hugging Face include such information.[70] The volume of data used for training AI systems has grown substantially in recent years, which has directly driven a corresponding rise in energy consumption across the AI lifecycle. With increased recognition into the link between data practices and environmental impacts[71], there are now targeted interventions at dataset management to reduce unnecessary data[72], alongside improvements in hardware and modelling efficiency to counteract the effects of more voluminous datasets.[73] A number of the model providers who were ranked as more mature in this aspect included details of the calculators and methodologies they used to document the environmental impact of the model, as well as details of mitigation strategies.

Although our approach was targeted at the two user groups that should be best catered for, the inconsistency of the results indicates substantial room for improvement. As we continue to apply the AIDTI methodology to new user groups and models, findings may change. Our expectation is that results will probably be less positive for some of the remaining user groups, for instance the public or media, which have largely been overlooked in current AI transparency discussions. For transparency information to be actionable, it is essential that model developers make the information available in a format that is accessible and useful to those groups.

[69] Gilardi, F., Alizadeh, M. and Kubli, M. (2023), 'ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks'.
[70] Liang, W. et al. (2024), 'Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI'.
[71] Wu, C.-J. et al. (2021), 'Sustainable AI: Environmental Implications, Challenges and Opportunities'.
[72] Verdecchia, R. et al. (2022), 'Data-Centric Green AI: An Exploratory Empirical Study'.
[73] Desislavov, R., Martínez-Plumed, F. and Hernández-Orallo, J. (2023), 'Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning'.

Discussion

Assessment method

We asked the assessors to provide feedback on how they found the process, their thoughts about the individual dimensions, and their recommendations for the next stages.

Element	Feedback from assessors
Process	At points, some assessors felt the maturity criteria placed too high expectations upon model developers to be able to document thoroughly. Future assessment could be used to identify the minimum viable approach for data transparency, to support developers to meet that standard. At the same time, assessors were encouraged to see evidence that most AI providers are at least aware of more than 50% of the assessment dimensions, with some making an effort to disclose information in a useful way.
What datasets are used throughout the lifecycle?	This was mostly straightforward for the assessors to do, although it was challenging to determine the extent to which this could be implemented for more complex models – such as those accessed in bulk, which should be used as a clear argument against the use of large, poorly documented datasets. A more granular approach to documentation, perhaps through using the AI data taxonomy[74], would be welcome here.
How was the data collected?	For this metric, the assessors requested a more granular approach and a stronger consideration of whether the data source – scraped, open source, enterprise – impacts how it should be documented. If synthetic data continues to be used in models, transparency over how this data was created and used will become vital.
What pre-processing activities took place?	This was mostly received positively, although some pointed out that some of these activities could be considered standard practice and therefore might not be documented that thoroughly. Conversely, some of the activities for large-volume models might require a significant resource outlay to achieve higher maturity. This fits the need to document data-centric benchmarks and evaluations such as ensuring the role of data in these earlier stages of the AI lifecycle is more clearly described – and to advocate for higher uptake of such approaches given that they are not yet common practice.[75]
Do they use an accessible mechanism for transparency?	There was a request to include points such as consistency and updates within the high maturity, to ensure all documentation is up to date. Likewise, model cards were more typically used than data cards, so incorporating both would improve the justification for this dimension.
Was copyrighted data used in the AI model/personal information in data	These dimensions received very similar feedback. They were mostly received well, although again there were some requests for further granularity in the assessment to consider the difference of when copyright data/PI might be included in the lifecycle. Even for the more complex models, there was a recognition that a simple disclosure of

[74] Hardinges, J and Simperl, E. (2024), 'A data for AI taxonomy'.
[75] Sambasivan, N. et al (2021), ''Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI'.

	whether copyright data/PI was included should be possible for all developers, particularly as the EU's transparency summary templates provide guidance on doing this.[76]
Environmental impacts of model	There was a recognition that this is only just beginning to be included within release notes, which means that older models were unlikely to include this information. Further, as there is no consistent approach to this documentation, comparison between models is difficult. Particularly as the dimension does perhaps not specify enough for which stage of the lifecycle we were looking for this information (i.e. just the training stage, or also operation?). Further work needs to be done to tie environmental impacts with data practices to justify the inclusion of this dimension in future assessments, alongside ensuring that model providers are aware of the needs for this data type.
Human/organisational supply chain	The assessors struggled the most with this dimension, with very little information regarding supply chains. Further, some of the global, collaborative supply chains are so complex that documenting them thoroughly is challenging. The assessors pointed out how much of the data work here falls under data pre-processing and data augmentation, both of which are often poorly documented aspects of the AI lifecycle.[77]
Suggestions for further dimensions:	It was suggested that we should consider developing open feedback mechanisms, such as whether the developers provide channels for users to report issues, biases, or harmful outputs, and document how this feedback is incorporated into future iterations.

Recommendations

To address the challenges, we have identified regarding data transparency across AI lifecycles during this project, we propose:

- **Building clarity about expectations of AI developers:** When considering the diverse needs for transparency information, it is clear that not all of these needs can simultaneously be addressed. There is a need to understand the key transparency 'sticking points' for model providers – i.e. why the numerous attempts to improve transparency practices do not achieve the desired results. This clarity can be established within policy that incentivises developers, but also through the development of a clearer 'gold standard' that accounts for diverse needs.
- **Governments, regulators and policymakers continue to drive up standards for minimum viable transparency and data documentation:** There will be a need to evaluate how the EU AI Act operates in the coming years, alongside the sufficiently detailed summary requirement.[78] As more regulators push for transparency, the approaches

[76] Warso, Z., Gahntz, M. and Keller, P. (2024), 'Sufficiently detailed? A proposal for implementing the AI Act's training data transparency requirement for GPAI'.
[77] Strasser, S, Klettke, M (2024), 'Transparent Data Preprocessing for Machine Learning'.
[78] Warso, Z, Gahntz, M and Keller, P (2024), 'Sufficiently detailed? A proposal for implementing the AI Act's training data transparency requirement for GPAI'.

in different jurisdictions will need to be consistent and somewhat standardised, as a too onerous burden on model providers could undermine thorough documentation attempts.

- **Further pushes for accountability:** There needs to be continued research and campaigning from researchers, journalists and policymakers to hold developers accountable, based on the information they are transparent with, and further pushes to increase the quality of the transparency information to work for the documented user needs.
- **Ensure transparency approaches are repeated for downstream use.** Our approach has centred upstream AI development, but transparency is also necessary about the deployment and monitoring of AI systems.[79] For example, access to model weights[80] and usage data[81] allows researchers to conduct further investigations to ensure AI safety.

Future work

Expanding the AIDTI

There are a number of developments that would improve the quality and relevance of the AIDTI, such as putting more models through the assessment. For example, this could be repeated for models within a particular jurisdiction, such as the EU, or an emphasis on models in different sectors or use cases.

Further validation of findings

The use of the maturity criteria ensured a subjective element to the assessment. Although the risks connected with subjectivity were somewhat mitigated via a validation process led by an ODI researcher to ensure that the assessments were carried out in accordance with our instructions, the results would be further validation if more than one assessor worked on a single model. This would allow the results to be cross-checked for accuracy and consistency. Similarly, designing mechanisms to enable users to conduct their own assessments of models, or provide feedback or suggestions to our assessments, would provide further validation.

Improving user-centricity

We have focused our design and research approach on developers and

[79] Hardinges, J. and Simperl, E. (2024), 'A data for AI taxonomy'.
[80] Open Source Initiative (n.d.), 'The Open Source AI Definition – 1.0'.
[81] Nicholas, G. (2024), 'Grounding AI Policy: Towards Researcher Access to AI Usage Data'.

RAI researchers in this instance, but we are looking to future opportunities to centre the design on the needs of further stakeholders beyond research environments, and target the needs of these groups. For example, a member of the public has different needs from a RAI researcher; it has been noted that 'not all users require an understanding of the 'depth' of AI, or all of the data used and explained, some need the most relevant and impactful information to be 'surfaced'[82] – meaning the information isn't only provided in lengthy technical reports.

Furthermore, for regulators, for users, and for deployers who may need more conveniently located information about particularly salient topics, research is needed to replicate our assessment, but through the lens of these other personas. A future version of the AIDTI, therefore, will include varying assessments and maturity scores based on how well they meet each persona's needs.

An interactive system, supported by machine-readable data

Firstly, a place or method for developers to collect and document the vital transparency information in a consistent way. This could build on established approaches – such as model cards – or it could be a new site or standard specifically for data transparency. Secondly, users (whether other developers, researchers, policymakers or the public) require this information in a form that can enable them to carry out their activities; the information therefore needs to be in a comparable and preferably machine-readable format.

In the future, we envision that the AIDTI will be hosted on a website that would become a consistent and singular repository for AI data transparency assessments. Further, this information would include the requested transparency data (such as lists of dataset sources) in a machine-readable format and we would explore its integration with existing environments such as Hugging Face.

Data lifecycle supply chain approach

There is also a growing need to be able to understand data across the AI lifecycle and recognise that the model developers are only one of the actors involved. A more advanced transparency system could allow a user to explore the full supply chain, and see organisations and their data practices, by building on approaches such as ecosystem graphs.[83] In practice, this could work in a similar manner to other supply chain transparency. For example, Open Supply Hub is a supply chain mapping platform that allows collaborative data sharing. Within this model,

[82] Cellard, L. (2022), 'Surfacing Algorithms: An Inventive Method for Accountability'.
[83] CRFM (n.d.). 'Ecosystem Graphs for Foundation Models'.

developers and technology companies could ‘claim’ their profile and ensure that the transparency information about their organisation is fair and accurate.

Limitations of approach

Our approach requires one or multiple assessors to scan published materials and apply a maturity score. The scalability of this approach, to consider more models, or more dimensions, is in doubt. Subjective analysis of the assessor, limited time for validating findings, and having no scope to check results with developers, are also factors. This was mitigated, to a degree, by our insistence that the transparency information had to be accessible; if it wasn’t easily found via search engines with obvious search terms, that is not accessible.

Due to the global state of the AI sector, and the range of organisations that signed up to the Seoul agreement, models have been developed in countries where English is not the first language. While we appreciate the ethical issues with the supposition that developers should publish in English in order to be considered ‘mature’ in our assessment, our selection criteria was to identify developers who have signed up to the Seoul agreement, which we suppose signals an intention to work with, or within, English-speaking markets. In addition to this, the assessment team included individuals with an understanding of French, Chinese and Korean.[84]

[84] There was no Arabic speaker for the models from the UAE, but the details were published in English anyway.

Annex A: User needs table

Persona	Example applications	Information needed
Data scientists and developers	• Looking to build or fork models • Choose which models to use/build on	• Technical information and limitations • How data was collected • Comparisons
RAI researchers	• Need to evaluate models • Design interventions grounded in emerging practices	• Technical information • The information to be presented constantly for comparison studies
Policymakers and regulators	• Ensure developers are following law and assessing impact on citizens • Understand existing practices to devise policy remedies	• Data collection methods • A detailed summary of the training content and data to meet regulatory requirements (such as the EU AI Act) • Data sources • Human/labour – to understand jurisdictions
Members of the public	• Choose ethical AI tools • Understand if data about them has been included in AI training	• Comparisons • Data sources • Environmental impacts
Creatives	• Understand if model has been trained on copyrighted data	• Data collection methods • Data sources
Journalists	• Compare and assess models and practices • Expose harmful data practices	• Data collection methods • Data sources • Environmental impacts
Lawyers	• Understand if personal or copyright data has been wrongfully included in training set	• Data collection methods • Data sources • Human/labour

Annex B: Models assessed

Model developer	Model name	Country	Frontier	OSAI
Amazon AWS	Amazon Titan Text Premier	USA	TRUE	FALSE
Anthropic	Claude 3.5 (Sonnet)	USA	TRUE	FALSE
BigScience	BloomZ	Global	FALSE	TRUE
Cohere	Aya	USA	TRUE	FALSE
Google	Gemini 1.5	USA	TRUE	FALSE
IBM	Granite	USA	TRUE	FALSE
Inflection AI	Inflection-2	USA	TRUE	FALSE
Meta	Llama 3.2	USA	TRUE	FALSE
Microsoft	Phi-3	USA	TRUE	FALSE
Mistral	8B	France	TRUE	FALSE
Naver Corporation	HyperCLOVA X	South Korea	TRUE	FALSE
OpenAI	GPT-4o	USA	TRUE	FALSE
Samsung	Gauss	South Korea	TRUE	FALSE
Stability.ai	Stable Diffusion 3.5	UK	FALSE	FALSE
Technology Innovation Institute (TII)	Falcon 2	UAE	TRUE	TRUE
xAI	Grok-2	USA	TRUE	FALSE
ZhipuAI & THUDM	GLM	China	TRUE	FALSE
Eleuther AI	Pythia	USA	FALSE	TRUE
BigCode	Starcoder2 15B	Global	FALSE	TRUE
LLM360	Amber	USA	FALSE	TRUE
AI2	OLMo-7B	USA	FALSE	TRUE
Inception / G42	JAIS 70B	UAE	TRUE	FALSE