EthicAmanuensis: supporting machine learning practitioners making and recording ethical decisions

1st Dave Murray-Rust
Industrial Design Engineering
Delft University of Technology
The Netherlands
d.s.murray-rust@tudelft.nl

2nd Konstantinos Tsiakas
Industrial Design Engineering
Delft University of Technology
The Netherlands
k.tsiakas@tudelft.nl

***Abstract*—Ethics should be a practice, not a checkbox. Data scientists want to answer questions about individuals and society using the vast torrent of data that flows around us. Machine learning practitioners want to develop and connect complex models of the world and use them safely in critical situations. Ethical issues can be seen as getting in the way of the core idea and form pain points around managing, using and learning from data, as well as designing human-centric and ethical systems. This is because there is a *design gap* around ethics in data science and machine learning: the tools that we use do not support ethical data use, which means that data scientists and machine learning practitioners, already engaged in technically complex, multidisciplinary work, must add another dimension to their thinking. This work proposes and outlines an infrastructure and framework that can support in-the-moment ethical decision making and recording, as well as post-hoc audits and ethical model deployment.**

***Index Terms*—Ethical AI; fairness; ethical annotations tool**

I. Introduction

Developing an ethical practice of machine learning alongside carrying out technical work demands a lot from practitioners and organisations, as "engineers of AI systems are increasingly expected to go beyond the traditions of requirement specifications, taking into account broader societal contexts and their complexities [20]". This has led to the emergence of *Ethical AI* as an active field, from concerns about the safety of ML systems [14] to the development of novel ways to understand computational ethics (e.g. [4]), the spread of ethics into computer science education [13] and so on. The scope of AI ethics is broad, not least because "in the past several years, seemingly every organization with a connection to technology policy has authored or endorsed a set of principles for AI" [9], with over 70 frameworks released as of 2020 [17].

While engineers may have been "doing ethics by other means; as they materialise moral decisions in the artefacts they create [24]", there remains space for technical mediation between high-level goals and the reality of daily practices. Practitioners must navigate between organisational constraints and demands, such as speed and performance and the need to create ethically defensible solutions. This is not easy, and "moving effectively from ethical theory and principles into context specific, actionable practice is proving a significant barrier for the widespread uptake of systematic ethical impact analysis in software engineering" [17].

One avenue to address this is the development of ways to better account for engineering decisions by making it part of the process - enlisting technology to *co-shape* human decisions in Verbeek's terminology [24]. This brings together several strands of thinking. Firstly, better accounts of decisions and processes support auditing and controls that help enforce principles (e.g. [8]). Secondly, there is a sense that there is a moral requirement to be able to justify and explain how an outcome was arrived at [6]. This must go beyond the functioning of an algorithm or model, and include information about how it was created and designed. Finally, Shklovski et al's work looks at how engineers can find 'nodes of certainty' when operating outside their narrowly defined technical practices and "identify the responsibilities that need to be in place to sustain trust and to hold the relevant parties to account" [20].

The aim of this paper is not to argue that all of ethics can be encoded computationally in the way that type systems or style guides can; rather, that there is a *significant* set of ethical concerns that can be meaningfully brought into the development process, and that doing so can create new sites and strategies for ethical discussion and behaviour. We propose the development of a tool that supports ethical AI/ML practices, with four key intentions:

- Support organisations in committing to ethical practices around data science and machine learning development.
- Support programmers during code implementation to consider ethical issues and record their decision making.
- Support system builders while combining data and models, considering the possible implications
- Support auditing and regulation of AI and ML driven systems during all stages of the ML/AI development

In order to do this, we develop a declarative specification that allows for the annotation of programs with ethical concerns and decisions. These annotations can be published without revealing exact algorithms, allowing public commitment to ethical principles without disclosing 'secret sauce' computational techniques. The aim of this framework is to prompt developers to engage with such decisions *in the moment*, and record the decision that has been made, so that it can propagate through the model process as a form of decision provenance [21].

II. Related Work

Ethical concerns and bias in AI/ML systems can arise during the several stages of the lifecycle of development, and in different forms. One of the existing frameworks for the classification and analysis of bias in computer systems identifies three main categories: *pre-existing*, *technical*, and *emergent* bias [10]. Focusing on data acquisition, collection and annotation, research aims to identify and address ethical challenges related to privacy, accountability and transparency issues. In order to ensure responsible and representative data acquisition and analysis, factors including *Auditability*, *Benchmarking*, *Confidence*, *Data-reliance*, and *Explainability* should be taken into consideration during the early planning of data acquisition and analysis [1].

With regard to the development and integration of an AI system in real-world applications, we can look at practices to improve ethical behaviours, e.g. *auditing*, *benchmarking*, *confidence and trust*, *explainability and interpretability* [2]. One can also focus on the sources of harm that arise through the ML lifecycle, for example Suresh's framework that covers aspects from data collection to model development and deployment [23], categorizing biases as: *Historical*, in *Representation*, *Measurement*, *Aggregation*, and the *Learning Evaluation* and *Deployment* bias. The goal of this framework is to provide a structure to understand possible problems and their sources and identify appropriate mitigation techniques, considering that there is not a generalizable set of solutions for the several possible problems during the ML lifecycle.

There is an importance to storing contextual information, metadata, and explanations as annotations during data collection and curation. *Fides* [22] implements a data sharing and collaborative analytics platform with features to promote best practices at all stages of the data science lifecycle. In terms of technical bias during ML pipelines, frameworks and tools have been proposed to identify and mitigate bias during model development and deployment [19]. *FairPrep* is a design and evaluation framework for fairness-enhancing interventions in machine learning pipelines during model development and deployment [18]. Focusing on concerns that may arise during data pre-processing, *fairDAGS* have been proposed for dataflow representation of ML pipelines to help identify possible concerns of bias [25].

III. Proposed Framework and Demonstration

The motivation of our proposed system is the need to support ethical annotations of the various stages of the ML lifecycle. Figure 1 gives an overview of the way that this model can be applied around the coding process. First, a discussion of high-level ethical concerns and behaviours leads to a specification for the ethical concerns that should be addressed during model development and deployment. This forms part of the context for the software development process, where a dataflow is developed and annotated to describe the models and data that it contains. Finally, the set of concerns can be checked against the annotated dataflow, to point out areas where concerns do not seem to have been met (warnings) and saved alongside model outputs to give a record of what has happened and document the ethical decisions made. The body of this section walks through key components of this process.

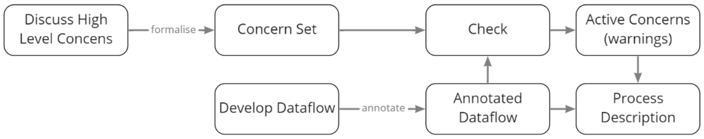


Fig. 1. Overview of system operation showing the process from discussion concerns, through formalising them into a concern set, annotating a dataflow and producing a description of the ethical aspects of the development process

A. Creating Concern Specifications

The first stage of the process (Figure 1) is to assemble a set of concerns that should be addressed. These come from a range of places - legal regulations and requirements such as the GDPR or the EU AI Act, organisational codes of conduct and data protection policies, public commitments to ethical behaviour, general ethical principles. However, a practical commitment to ethical behaviour would require an agreement between system commissioners and system developers about what the ethical commitments should be, just as the performance of the system or its delivery time would be specified, and frameworks for carrying out this kind of process are emerging [26]. Therefore, we look at the creation of a 'concern specification' as both a publishable statement of intent and a declarative object that supports semi-automated checking of ethical concerns.

To make this more concrete, we use a set of example concerns that illustrate different stages of the data and modelling lifecycle. The idea is that these hard-edged checks represent a dialogue about what the top level values of the organisational process are, and how to translate them into actionable specifications:

- noPII - models and datasets should not be exported which expose personally identifying information to the outside world. As with many of this issues, this requires some sensitivity: a formal data provenance model might can prove some properties that are or are not exposed, but questions about e.g. whether PII can be extracted from a model are more contextual.
- allDataLabelled(pii ,protected) - an organisational specification that all datasets, fields and models need to label the data coming in and out as to whether it is PII or speaks to a protected class. This is a background concern that supports the rest, but is a necessary part of general data protection practices.
- balanced - all datasets used to train models should be have enough examples of each class to allow proper training. Datasets that contain few examples of particular outcomes or particular groups of the population are more likely to produce biased models, which can be addressed through procedures such as *fair-SMOTE* [5]

$$
\begin{aligned}
Field &:=< id, attrs > \\
Dataset &:=< id, Field*, attrs > \\
Model &:=< Dataset, Dataset, attrs > \\
Deployment &:= (Field|Dataset|Model)* \\
Artefact &:= Field|Dataset|Model|Deployment \\
\\
Assertion &:=< claim, Artefact*, attrs > \\
Domain &:= Artefact \rightarrow bool \\
Check &:=< Artefact, Assertion* > \rightarrow bool \\
Concern &:=< name, Domain, > \\
Instantiation &:=< Artefact, Check >
\end{aligned}
$$

Fig. 2. Annotation model for describing dataflow artefacts, along with claims made about their properties to be checked for ethical concerns.

- noProtectedInModels - Models should not be trained on fields that are protected attributes. This is a design choice made by the organisation in pursuit of both fairness and data security.
- fairness - Any models produced should not give discriminatory outcomes around protected attributes. There are many ways to check fairness, and multiple viewpoints [3] – we can assume that the organisation has chosen a group fairness metric, where subsets of the input that differ on protected attributes only do not have different outcomes.

B. Workflow Objects and Annotations

To provide a foundation for ethical annotation, we provide a simple model for annotations over elements of a dataflow (Figure 2). This model is not intended to model every aspect of the objects under discussion, but rather to have just enough structure to support meaningful annotation and support the possibility of automatic reasoning. We look at the following basic elements that describe the objects in a standard data science workflow:

- *Fields* as single columns in a Dataset (or similar collections of data with a consistent type).
- *Datasets* as being composed of a collection of fields;
- *Models* as mappings from a Dataset to an output Dataset
- A *Deployment* is a collection of these artefacts along with some kind of context
- All of these are *Artefacts* to be discussed.
- A *Reference* is used to point to another *Artefact* in order to make use of its properties.

Each of these can have extra information appended ($attrs$) consisting of key/value pairs to describe other features of the object. Objects can also be included by *reference* for conciseness.

C. Assertions and Concerns

We now introduce the notion of *assertions* and *concerns*, that relate the model description to the agreed on set of ethical concerns. This roughly follows the idea of argumentation theory, e.g., [16], that claims can be made about the state of the world, which should be backed up with some form of evidence. We also follow along with the intuition that when arguments are made over formal terms – in this case our universe of model annotations – the structure of these terms can be used to aid the reasoning [7].

Assertions (A) are positive claims for properties of fields, datasets, models, as well as direct answers to concerns. Each annotation is the name of a predicate, a specification of the artefacts to which it applies (target) and an open set of extra information about how this has been achieved, e.g. {claim:'balanced',domain:['d1.age','d1.income'],how:'smote'}. [1]

Concerns (C) represent the ethical concerns outlined in the concern specification (Section III-A). These are abstract specifications of concerns: they give an indication of where the concern *might* obtain – is it related to particular fields in the data, or only to a trained model? – and specify what kinds of claim would be needed to satisfy the concern. Each provided concern must specify:

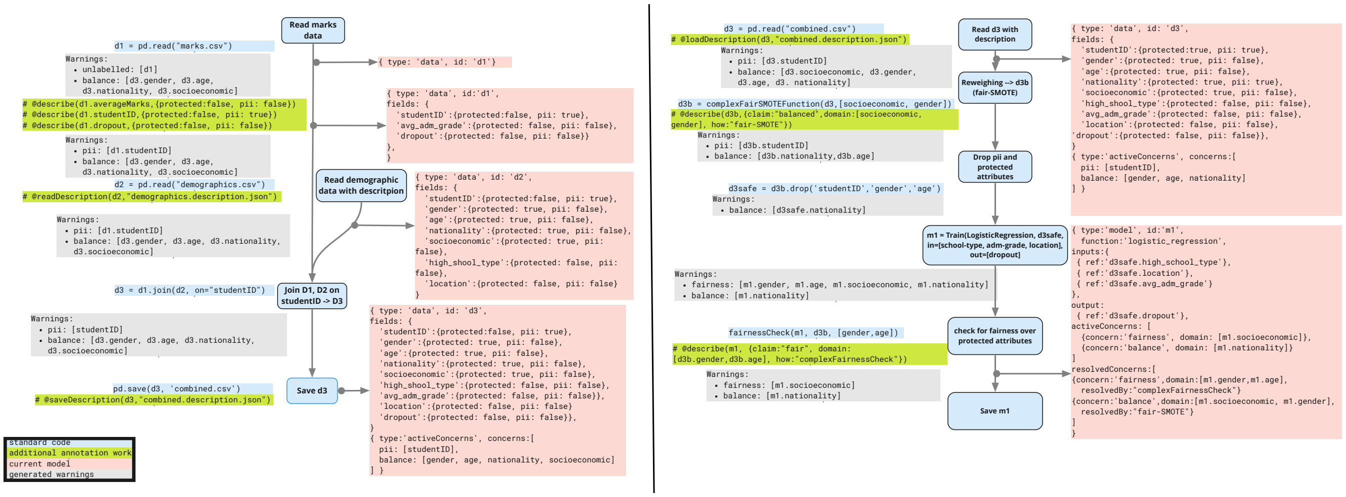
- The *domain* of the concern: the kinds of artefact to which it applies (models, datasets, fields or deployments) and prerequisites for the condition to apply e.g. 'this concern only applies to fields/models that contain personally identifying information.'.
- A *checking* process that specifies whether the concern has been answered by a particular set of assertions. The simplest form would be one that requires a certain assertion to be present, e.g. a concern about fairness might require a model to marked as fair in relation to all fields that are listed as protected attributes.

For some concerns, little technical specification can be given - the degenerate case is that a particular concern has to be explicitly answered by a particular claim, such as a fairness claim requiring a particular annotation of fairness. More complex processes could include sets of allowable annotations (e.g. for different fairness definitions), annotations on precursor objects (e.g. models are trained on balanced data if there is a claim on the dataset they are trained on that it is balance) and so on.

Instantiated concerns (I): given a set of artefacts and a set of assertions over them, each concern $c \in C$ will give rise to particular instantiations about whether the concern applies to particular artefacts, e.g., does *this* field need to be labelled? – does *this* model need to be checked for fairness, etc. Each of these instantiations can then be used to parameterise the concern's decision process for the point of application - the checking function can decide whether the concern has been answered for each artefact where it applies.

At this point, we have annotated the workflow to describe the models and data that are brought in and used, and we

[1] as shorthand within the JSON notation, we assume that: (a) the target of a nested assertion is its parent object, i.e., adding {claim:'pii'} to the body of the field d1.age is equivalent to {claim:'pii', target:['d1.age']} as a single object, and (b) attributes that are the names of claims can be taken as asserting that claim, so { field:"age", protected:"true", pii: "false"} is equivalent to { field:"age", claims: [{ claim:'protected'}, { claim: 'nopii' }] }


Read marks data
Read demographic data with descritpion
Join D1, D2 on studentID -> D3
Save d3
standard code
additional annotation work
current model
generated warnings
Read d3 with description
Reweighing --> d3b (fair-SMOTE)
Drop pii and protected attributes
m1 = Train(LogisticRegression, d3safe, in=[school-type, adm-grade, location], out=[dropout]
check for fairness over protected attributes
Save m1

```
d1 = pd.read("marks.csv")
Warnings:
  • unlabelled: [d1]
  • balance: [d3.gender, d3.age,
    d3.nationality, d3.socioeconomic]
# @describe(d1.averageMarks,{protected:false, pii: false})
# @describe(d1.studentID,{protected:false, pii: true})
# @describe(d1.dropout,{protected:false, pii: false})
Warnings:
  • pii: [d1.studentID]
  • balance: [d3.gender, d3.age,
    d3.nationality, d3.socioeconomic]
d2 = pd.read("demographics.csv")
# @readDescription(d2,"demographics.description.json")
Warnings:
  • pii: [d1.studentID]
  • balance: [d3.gender, d3.age, d3.nationality,
    d3.socioeconomic]
d3 = d1.join(d2, on="studentID")
Warnings:
  • pii: [studentID]
  • balance: [d3.gender, d3.age, d3.nationality,
    d3.socioeconomic]
pd.save(d3, 'combined.csv')
# @saveDescription(d3,"combined.description.json")
```

```
{ type: 'data', id: 'd1'}
```

```
{ type: 'data', id:'d1',
fields: {
  'studentID':{protected:false, pii: true},
  'avg_adm_grade':{protected: false, pii: false},
  'dropout':{protected: false, pii: false}}
},
}
```

```
{ type: 'data', id: 'd2',
fields: {
  'studentID':{protected:false, pii: true},
  'gender':{protected: true, pii: false},
  'age':{protected: true, pii: false},
  'nationality':{protected: true, pii: false},
  'socioeconomic':{protected: true, pii:
false},
  'high_shool_type':{protected: false, pii:
false},
  'location':{protected: false, pii: false}
}
```

```
{ type: 'data', id: 'd3',
fields: {
  'studentID':{protected:false, pii: true},
  'gender':{protected: true, pii: false},
  'age':{protected: true, pii: false},
  'nationality':{protected: true, pii: false},
  'socioeconomic':{protected: true, pii: false},
  'high_shool_type':{protected: false, pii: false},
  'avg_adm_grade':{protected: false, pii: false}},
  'location':{protected: false, pii: false}
  'dropout':{protected: false, pii: false}},
}
{ type:'activeConcerns', concerns:[
  pii: [studentID],
  balance: [gender, age, nationality, socioeconomic]
] }
```

```
d3 = pd.read("combined.csv")
# @loadDescription(d3,"combined.description.json")
Warnings:
  • pii: [d3.studentID]
  • balance: [d3.socioeconomic, d3.gender,
    d3.age, d3. nationality]
d3b = complexFairSMOTEFunction(d3,[socioeconomic, gender])
# @describe(d3b,{claim:"balanced",domain:[socioeconomic,
gender], how:"fair-SMOTE"})
Warnings:
  • pii: [d3b.studentID]
  • balance: [d3b.nationality,d3b.age]
d3safe = d3b.drop('studentID','gender','age')
Warnings:
  • balance: [d3safe.nationality]
Warnings:
  • fairness: [m1.gender, m1.age, m1.socioeconomic, m1.nationality]
  • balance: [m1.nationality]
fairnessCheck(m1, d3b, [gender,age])
# @describe(m1, {claim:"fair", domain:
[d3b.gender,d3b.age], how:"complexFairnessCheck"})
Warnings:
  • fairness: [m1.socioeconomic]
  • balance: [m1.nationality]
```

```
{ type: 'data', id: 'd3',
fields: {
  'studentID':{protected:true, pii: true},
  'gender':{protected: true, pii: false},
  'age':{protected: true, pii: false},
  'nationality':{protected: true, pii: true},
  'socioeconomic':{protected: true, pii: false},
  'high_shool_type':{protected: false, pii: false},
  'avg_adm_grade':{protected: false, pii: false},
  'location':{protected: false, pii: false},
'dropout':{protected: false, pii: false}},
}
{ type:'activeConcerns', concerns:[
  pii: [studentID],
  balance: [gender, age, nationality]
] }
```

```
{ type:'model', id:'m1',
  function:'logistic_regression',
inputs:{
 { ref:'d3safe.high_school_type'},
 { ref:'d3safe.location'},
 { ref:'d3safe.avg_adm_grade'}
},
output:
 { ref:'d3safe.dropout'},
activeConcerns: [
  {concern:'fairness', domain: [m1.socioeconomic]},
  {concern:'balance', domain: [m1.nationality]}
]
resolvedConcerns:[
{concern:'fairness',domain:[m1.gender,m1.age],
  resolvedBy:"complexFairnessCheck"}
{concern:'balance',domain:[m1.socioeconomic, m1.gender],
  resolvedBy:"fair-SMOTE"}
]
}
```

Fig. 3. Running examples. Left - Example workflow for combining a dataset containing student demographic information with one that contains marks and outcomes from their activity on the course. Right - Workflow for training a model over the combined dataset: i) read in a dataset, ii) address a warning about a lack of balance through a re-balancing procedure, iii) address a warning about personally identifying information through dropping the relevant column, iv) train a model on the 'safe' dataset, v) checking for fairness and recording that this has been done, vi) save the model along with its description.

have specified a set of concerns that may be raised through the development process. This has given rise to a set of instantiations - places where the concern applies, and may or may not have been answers. This set is the key output of the process. It provides firstly a record of the process in terms of addressing key ethical issues. It also provides the potential for an editor to raise all of the ethical concerns as warnings or errors – every instantiation that is not properly addressed is a warning to the developer. This gives programmers the possibility to:

- ignore an instantiation of a concern, and have it be present in the final output
- introduce a procedure that ameliorates the concern
- provide a reason why the concern is not relevant to this particular process, and record that reason.

This final structure, with a set of active concerns, the list of concerns that have been addressed and the ways in which they have been addressed, and the descriptions of important parts of the model can then be saved for future inspection.

D. Demonstration with worked example

Out of the many possible fields to examine for motivating examples, we define our problem following the example from [15], and look at the creation of a model that predicts whether students are likely to drop out of university. There are concerns with this kind of modelling over the representativeness of data, and the fairness of the models outcomes, but also the ways that those outcomes are used and the actions taken on them. In this example, we look at the following process:

- A background database of student demographics is combined with a database of performance, including whether the student drops out. The background data includes protected characteristics (age, gender, socioeconomic background, nationality) that should not be shared or used in decision making, but need to be present for questions around fairness.
- After combination, the data goes through pre-processing phases and is used to train a model from demographics to a likelihood of dropping out.
- Tests are brought in for certain kinds of fairness over the resulting model, which is then deployed in some decision making capacity.

We illustrate the annotation process of the workflow for our worked example. We present the different stages of the pipeline; data acquisition (load and combine datasets), data pre-processing, and model training (Figure 3). We visualize the 'base' data science code (in blue) along with the additional annotations that would be required (in green), the current description of the artefacts (in red) and the warnings that would be produced given the example set of concerns (in grey). The set of concerns from III-A is used, specifying that all data should be labelled as to whether it is personally identifying or pertains to protected attributes; no PII should be exported in models or datasets, all datasets used in models should be balanced with respect to protected attributes, and models should not discriminate between groups that only differ on protected attributes.

IV. Conclusion

In this paper, we propose a framework for representing and carrying out simple reasoning over data science and machine learning workflows in relation to ethical concerns. As noted previously, we see this tool as part of an ethical ecosystem (Figure 4). Previous work, in particular FairDAGs [25] and mlinspect [11], [12] have shown that it is possible to extract graph-based representations of naturally created machine learning pipelines that are amenable to investigation. This paper proposes a layer that mediates between these low-level structures and high-level organisational concerns

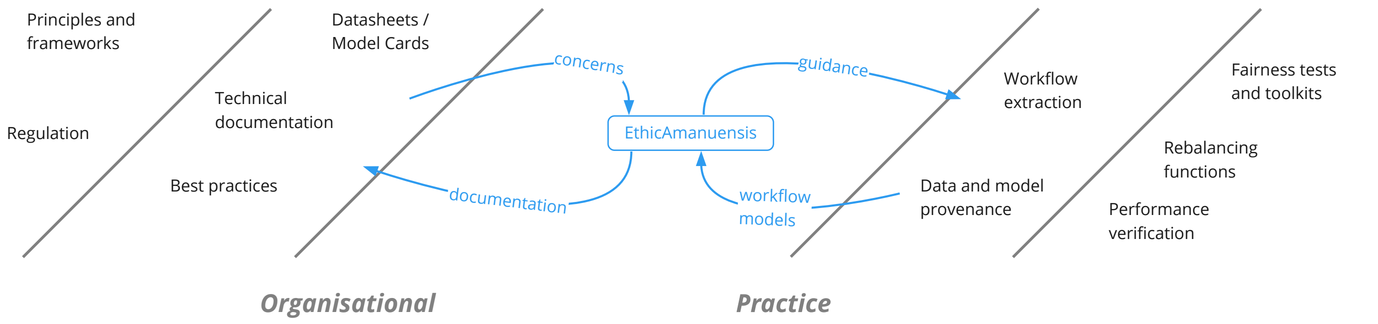


Fig. 4. Illustration of the position of the current work in a data ethics ecosystem – helping to codify high level concerns from the organisational perspective and guiding the application of emerging ethical machine learning practices, then drawing on detailed representations to build up a comprehensive documentation of decision provenance around ML development.

around ethical behaviour. The intent is that this process is to collectively specify concerns, taking them into the development process at the same conceptual level as existing warnings and errors provided by coding environments, and then documenting the decisions made. We hope this will support ethical processes around model building, in particular the envisioning of potential impacts before they happen and the performance of algorithmic impact assessments. While this is a model description paper and does not provide a concrete implementation, we hope that we have illustrated an important approach – between the computationally implementable approaches to tracking data through models and the higher level discourses around ethical frameworks lies a space that can be navigated by simple, semi-formal approaches that use as much computational support as is reasonable to support human and organisational decision making, recording and communication around ethical practices.

References

[1] Alice Baird, Simone Hantke, and Björn Schuller. Responsible and representative multimodal data acquisition and analysis: on auditability, benchmarking, confidence, data-reliance & explainability. *arXiv preprint arXiv:1903.07171*, 2019.

[2] Alice Baird and Björn Schuller. Considerations for a more ethical approach to data in ai: on data representation and infrastructure. *Frontiers in big Data*, 3:25, 2020.

[3] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 514–524, 2020.

[4] Abeba Birhane. Algorithmic injustice: A relational ethics approach. 2(2):100205.

[5] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.

[6] Mark Coeckelbergh. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. 26(4):2051–2068.

[7] Joseph Corneli, Ursula Martin, Dave Murray-Rust, Gabriela Rino Nesin, and Alison Pease. Argumentation theory for mathematical argument. *Argumentation*, pages 1–42, January 2019.

[8] Ray Eitel-Porter. Beyond the promise: Implementing ethical AI. 1(1):73–80.

[9] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. (2020-1).

[10] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

[11] Stefan Grafberger, Shubha Guha, Julia Stoyanovich, and Sebastian Schelter. Mlinspect: A data distribution debugger for machine learning pipelines. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2736–2739, 2021.

[12] Stefan Grafberger, Julia Stoyanovich, and Sebastian Schelter. Lightweight inspection of data preprocessing in native machine learning pipelines. In *Conference on Innovative Data Systems Research (CIDR)*, 2021.

[13] Barbara J. Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. Embedded EthiCS: Integrating ethics across CS education. 62(8):54–61.

[14] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved Problems in ML Safety.

[15] Marzieh Karimi-Haghighi, Carlos Castillo, Davinia Hernandez-Leo, and Veronica Moreno Oliver. Predicting Early Dropout: Calibration and Algorithmic Fairness Considerations.

[16] Hugo Mercier and Dan Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.

[17] Dorian Peters, Karina Vold, Diana Robinson, and Rafael A. Calvo. Responsible AI—Two Frameworks for Ethical Design Practice. 1(1):34–47.

[18] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In *Proceedings of the 23nd International Conference on Extending Database Technology, EDBT, 2020*, 2020.

[19] Sebastian Schelter and Julia Stoyanovich. Taming technical bias in machine learning pipelines. *Bulletin of the Technical Committee on Data Engineering*, 43(4), 2020.

[20] Irina Shklovski and Carolina Némethy. Nodes of certainty and spaces for doubt in AI ethics for engineers. 0(0):1–17.

[21] Jatinder Singh, Jennifer Cobbe, and Chris Norval. Decision Provenance: Harnessing data flow for accountable systems. 7:6562–6574.

[22] Julia Stoyanovich, Bill Howe, Serge Abiteboul, Gerome Miklau, Arnaud Sahuguet, and Gerhard Weikum. Fides: Towards a platform for responsible data science. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–6, 2017.

[23] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9. 2021.

[24] Peter-Paul Verbeek. Materializing morality: Design ethics and technological mediation. 31(3):361–380.

[25] Ke Yang, Biao Huang, Julia Stoyanovich, and Sebastian Schelter. Fairness-Aware Instrumentation of Preprocessing˜Pipelines for Machine Learning.

[26] Mireia Yurrita, Balayn Agate, Dave Murray-Rust, and Alessandro Bozzon. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. 2022.