AI and Ethics
https://doi.org/10.1007/s43681-024-00535-1

ORIGINAL RESEARCH

Capturing the unobservable in AI development: proposal to account for AI developer practices with ethnographic audit trails (EATs)

Yung-Hsuan Wu[1]

Received: 21 June 2024 / Accepted: 20 July 2024

Abstract
The prevalence of artificial intelligence (AI) tools has inspired social studies researchers, ethicists, and policymakers to seriously examine AI's sociopolitical and ethical impacts. AI ethics literature provides guidance on which ethical principles to implement via AI governance; AI auditing literature, especially ethics-based auditing (EBA), suggests methods to verify if such principles are respected in AI model development and deployment. As much as EBA methods are abundant, I argue that most currently take a *top-down* and *post-hoc* approach to AI model development: Existing EBA methods mostly assume a preset of high-level, abstract principles that can be applied universally across contexts; meanwhile, current EBA is only conducted after the development or deployment of AI models. Taken together, these methods do not sufficiently capture the very developmental practices surrounding the constitution of AI models on a day-to-day basis. What goes on in an AI development space and the very developers whose hands write codes, assemble datasets, and design model architectures remain unobserved and, therefore, uncontested. I attempt to address this lack of documentation on AI developers' day-to-day practices by conducting an ethnographic "AI lab study" (termed by Florian Jaton), demonstrating just how much context and empirical data can be excavated to support a whole-picture evaluation of AI models' sociopolitical and ethical impacts. I then propose a new method to be added to the arsenal of EBA: Ethnographic audit trails (EATs), which take a *bottom-up* and *in-progress* approach to AI model development, capturing the previously unobservable developer practices.

Keywords AI ethics · Ethics-based auditing · Principle-based AI governance · Ethnography · AI audit trails · AI lab studies

1 Introduction

"If you want to understand the big issues, you need to understand the everyday practices that constitute them." [1].

Artificial intelligence (AI) models'[1] sociopolitical impacts have been widely recognized, especially given their gradual integration into various domains like healthcare, legal systems, insurance industries, etc. [2–6]. There have long been alarms sounding about the sociopolitical problems and ethical consequences these AI models engendered: Ivana Bartoletti challenged the perceived neutrality of data and exposed how deep-running real-world biases become modeled into machines in *An Artificial Revolution* [7]. Kate Crawford echoes the same sentiment, demonstrating in her *Atlas of AI* how the power dynamics and socioeconomic forces underlie the labor, data, classification systems, and outputs that lead to the creation of AI [8]. The field of AI ethics examines the sociopolitical and ethical issues that arise from AI models making significant decisions about humans and, therefore, advocate vehemently for a normatively grounded governance; an ideal form of AI governance should be based on a particular set of normative or ethical principles [9–12]. However, the promotion of ethical principles themselves is insufficient to ensure that they are respected throughout the development of AI models.

[1] In this paper, the term "AI model" refers to an algorithmic model that a computer builds partially without human intervention after observing some data and recognizing patterns from such data. The term "AI system" is used to describe an overall system consisting of multiple AI models.

✉ Yung-Hsuan Wu
yung-hsuan.wu@graduateinstitute.ch

[1] Graduate Institute of International and Development Studies, Geneva, Switzerland

Published online: 02 September 2024

The rapidly growing multidisciplinary field of ethics-based auditing (EBA) is precisely dedicated to evaluating and verifying if ethical principles are implemented throughout AI development. EBA provides "a structured process whereby an entity's present or past behavior is assessed for consistency with relevant principles or norms" [13–14]; it does so by translating abstract principles into "verifiable claims," which are "statements for which evidence and arguments can be brought to bear on the likelihood of those claims being true." [13]. EBA holds the promise of promoting procedural regularity, institutional trust, and transparency [15–17]. Not only has EBA garnered significant academic attention [15], but it has also piqued the interest of policymakers [18] and professional services firms [19–21]; a nascent yet headline-grabbing AI auditing industry is also emerging [22–25].

In this article, I argue that EBA's existing methods rely on applying highly abstract ethical principles to specific cases in a *top-down* and *post-hoc* approach, which does not always allow the complete evaluation of the sociotechnical assemblage of AI models (Sect. 2). In fact, current EBA methods tend to focus more on the AI models themselves as if they are distinct from the developers' practices and design processes that brought them into being in the first place; existing EBA methods then further hide the developers behind the curtain of *mechanical objectivity* and shielding them from scrutiny (Sect. 3). Using an ethnography framed as an "AI laboratory (lab) study," whose methods are explained in Sect. 4, I expose how the creation of AI models is a value-laden and sociopolitical process (Sect. 5). I then propose a *bottom-up* and *in-progress* approach termed an ethnographic audit trail (EAT) to reveal the ethical weights stemming from developer practices and design processes (Sect. 6). In Sect. 7, I conclude with two identified challenges for keeping EATs and offer suggestions for future research to substantiate the proposed method.

2 Ethics-based auditing (EBA)

2.1 Current methods

The litany of abstract principles alone—whether grounded in ethics, human-centeredness, or other normative systems—provides neither sufficient guidance for AI developers to implement them nor concrete evaluation frameworks that could be used to verify AI developers' self-proclaiming ethical practices. EBA comes into help: A variety of EBA procedures have been developed. The methods of translating abstract principles into verifiable claims are roughly divided along the lines of quantitative and qualitative procedures. The quantitative method relies on *metricizing abstract values*, constructing mathematical definitions of principles to describe the characteristics of training datasets and model performance [15]. The emerging field of Fair ML, for example, contributes a proliferating pool of tools that calculate fairness and measure algorithmic biases [26–27].[2] Other principles like transparency and accountability are similarly quantified [15].

The qualitative method relies on *operationalizing abstract values into verifiable statements*. Consider AstraZeneca's audit case [28]: The biopharmaceutical company hired an external auditor to examine both the high-level organizational structures and in-depth processes of its specific AI projects against its own published set of ethical principles [29]; The auditor took each principle and operationalized it into subsidiary and verifiable statements. For example, the principle of "Fair" was turned into two statements: "We endeavour to use robust, inclusive datasets in our Data [and] AI systems;" "We treat people and communities fairly and equitably in the design, process, and outcome distribution of our AI systems." [28]. The auditor then collected data from interviews and company documentation to corroborate the boiled-down claims.

2.2 Applying a critical lens on ethics-based auditing

As much as EBA has gained traction as a way of assessing the ethical implications of AI models, I argue that the existing methods of EBA are insufficient to capture all the ethical implications that might arise from the specific contexts in which the AI models are born and function. Quantitatively metricizing and qualitatively operationalizing abstract values take a *top-down* approach to evaluate the ethical implications of AI models, assuming a preset of abstract ethical values that are universally relevant and applicable to different instances of AI models across contexts. For example, one significant draw of metricizing ethical values is that by rendering the latter mathematical, one can apply the metrics across the board as if they are but a technical standard neutral from humans' subjective interpretations. This approach has been criticized for its "principlism" and "technical focus," as it directly applies abstract concepts across models built for a range of complex contexts without considering the particularities and contingencies of reality [16, 30–31]. On the other hand, qualitative operationalization of abstract values takes a similar *top-down* position; just as in AstraZeneca's case, the auditor must start with a decided preset of values to operationalize [28]. This may lead the auditor to leave out information not readily understood in

[2] Prominent Fair ML toolboxes include FAIRVIS, Microsoft Fairlearn, Google People and AI Research (PAIR)'s What-If Tools, IBM AI Fairness 360, University of Chicago Aequitas, etc.

the conceptual framework defined by the preset and fail to capture new and emerging ethical risks.

Moreover, I argue that current EBA methods miss an essential piece of the puzzle—the developmental process of the AI model and AI developers' specific practices that make the models capable of leaving any ethical impacts on the world in the first place. The current approaches in EBA are *post-hoc* in that ethical evaluations only occur after the model has been built, not before or during. Ethical metrics are often used to test *model performance*; in other words, they are only used to evaluate the model's capability to carry out ethical effects but not how they become capable. Reading between the lines of AI companies and researchers' documentation on the often metric-based performance tests, they mainly concentrate on explaining *the capacity of their models* instead of detailing how the latter acquired such capabilities [32–34].

On the other hand, the qualitative operationalization of abstract values is devised for audits that look beyond the technical and examine the influences of governance structures, managerial decisionmaking, and documentation of design choices [35]. While holding the promise to capture the broader sociopolitical context of an AI model, this type of audit is often carried out via interviews and official documentation after the model has been developed and even deployed. This means that the auditor also misses out on many contextual details in which ethical risks arise, such as what competing alternatives existed for a specific ethically contested design choice and what factors determined the final decision.

Taken together, the current *top-down* and *post-hoc* approach in EBA inadvertently masked the critical process by which an AI model is developed, making the developers' specific practices invisible. This is potentially an effect of the *mechanical objectivity* commonly associated with computer science and machine learning (ML) fields: The act of day-to-day coding and modeling is perceived to be purely based on practical functionality instead of sociopolitical considerations; hence, they do not need to be examined [36–37].

This mechanical objectivity of developers' practices can lead to unintended, negative consequences. Ugwudike showed that the theoretical foundations of developers underpin their design logic, which then affects how predictive policing algorithms operate and generate real-world effects [38]. Marino also demonstrated that codes are more than functional instructions for machines but also personal expressions of the programmers [36]. From conceptualizing a real-world problem to be solved, coding instructions into scripts, assembling datasets that become ground truths to the machines, designing a model architecture, and devising reward functions to selecting performance metrics, AI developers' implicit values and respective worldviews are embedded in the model every step of the way. AI models should not be considered immune from the sociopolitical influences of the larger world introduced by the hands of those who build them [39].

3 Discovering the new observable through ethnographic methods

There is a need to account for the new observable—the developer's practices that mix in their values and worldviews into AI models—to reveal developers' accountability by peeling away the façade of mechanical objectivity and to fully capture the root causes of ethical risks emanating from AI models. However, developers' theoretical foundations and implicit worldviews are often not readily quantifiable and describable by metrics [38]. Qualitative methods like conducting interviews and reviewing developers' documentation *post-hoc* are also insufficient; science and technology studies (STS) literature noticed that scientific reports and documentation tend to include only purified accounts of the decisions taken and provide step-by-step maxims of conduct in research or experimenting activities that discard and hide the scaffolding utilized to arrive at the scientific facts [40].

How can we capture this new observable? As informed by the STS literature, a non-top-down, non-post-hoc approach that allows us to examine developers' practices and unearth values embedded in AI models might be ethnographies and, more precisely, "AI laboratory (lab) studies," as proposed by Florian Jaton [40].

The justifications for an ethnographic method start with the theoretical reconceptualization of AI models as a sociotechnical assemblage: As Seaver observed, "algorithmic systems are not standalone little boxes, but massive, networked ones with hundreds of hands reaching into them, tweaking and tuning, swapping out parts and experimenting with new arrangements." [41]. AI models "must be understood as composites of nonhuman (i.e., technological) actors woven together with human actors, such as data-creators, maintainers, and operators into complex sociotechnical assemblages." [42]. Seen in this way, AI models are actively enacted by the practices of a myriad of actors that act on both technical and non-technical concerns [43]. The "intersection of dozens of…social and material practices" [44] that created AI models cannot be divorced from the broader contexts; an AI model must be understood as "relational, contingent, contextual in nature" instead of "technical, objective, impartial." [37, 45].

To unpack this complete sociotechnical assemblage, Kitchin proposed a combination of interviews, ethnographies, and document analyses, accounting for the

"infrastructure/hardware, code platforms, data and interfaces" that are framed and conditioned by "forms of knowledge, legalities, governmentalities, institutions, marketplaces, finance and so on." [37]. Consequentially, the venues for observation expand beyond the interview room where the managers and developers are arbitrarily taken outside of the scenarios where they practice and make decisions; instead, an observer must enter the exact places where AI models are under development and investigate the AI lab, the C-suite boardrooms, the cross-department meetings, developers' desks, coding scripts, datasets, application interfaces, data contractors, front-end engineers, company competitors, market forces, and users.

By documenting the developers as they work and interact with technical and non-technical components, an ethnography is fitting to witness the "everyday practices that constitute [the algorithms] and keep them working and changing." [43]. Jaton further considered a traditional analytical genre within STS called "laboratory studies." [40, 46–47]. Instead of starting from established scientific facts, lab studies concentrate on the "mundane actions and work practices to document and make visible how scientific facts were progressively assembled." [40].

By definition, an AI lab study takes a *bottom-up* and *in-progress* approach to examining the development of AI models. It doesn't start with a priori assumptions about what goes on in an AI lab and which set of values the ethnographer must pay attention to when documenting and problematizing practices, giving the ethnographer the total flexibility to note down any details. Moreover, as opposed to retrospectively examining a developed and deployed model, a lab study documents activities that occur *in progress* as "a set of intertwining courses of actions [which are accountable chronological sequences of gestures, looks, speeches, movements, and interactions among humans and nonhumans] sharing common finalities [such as ending up as a mathematical model, code, algorithm, or program]." [40].

In this paper, I apply AI lab study methods to a case to illustrate just how ethnographic methods can help capture previously unobservable developer practices. Especially bringing to the forefront the social practices developers engage in that influence the material practices, I will peel away the façade of mechanical objectivity of developer practices, showcase how developers encode subjective worldviews and social relations into AI models, and finally demonstrate why ethnographic methods to capture the new observable are critical in EBA and other ethical evaluations of AI models.

4 Methodology

There is no unified modus operandi of laboratory studies; there exist various viewpoints and approaches in the most renowned works [46, 48–50]. Nevertheless, a few threads run through most studies. Taking a constructionist approach, lab studies examine scientific activities via direct participant observation that generates detailed, thick descriptions; ethnographers then use discourse analyses to make sense of the themes and related components underlying the dense qualitative data [47]. During a lab study, an ethnographer documents the "technical activities of science within the wider context of equipment and symbolic practices," which treats the former as cultural activities [47]. In the same vein, technical objects are not to be considered "technically manufactured in laboratories" but "symbolically and politically construed." [47].

The practical steps of my lab study are straightforward: I located an AI lab where I could gain sufficient access to both the AI developers and the technical artifacts they work on; took notes during a variety of courses of action within the lab; followed specific processes that were parts of bigger projects; partook in meetings; conducted interviews to clarify facts; and analyzed findings.

The single-case study occurred in an AI lab within a Swiss-Maltese non-governmental organization (hereinafter "the Foundation"). The Foundation conducts capacity development activities supporting small and developing states in diplomacy, particularly in internet governance and digital policy. Apart from conducting research on policy processes and training, the Foundation also develops in-house technological products to test how various digital and, particularly, AI technologies could help diplomats' day-to-day work. I zoomed in on one of the Foundation's AI projects, which involves building an AI reporting system that generates just-in-time reports from international conferences and events.[3]

The data collection ran from October to December 2023. There were several sites for observation, primarily meetings with a different mix of people, one-on-one chats, semi-structured interviews, and self-explorations. The inquiry was hybrid in that I conducted in-person observation via two field trips to Belgrade, Serbia, where the AI lab is, and partook in online processes like team and brainstorming meetings. I collected audio and video recordings,[4] observation notes,[5] drawn illustrations, and text documents. I also

[3] I refer to a specific occasion of international conferences and events as "events"; there are usually multiple "sessions" at an "event."

[4] Audio and video recordings were obtained from meetings and interviews with participants' consent to record. They were further transcribed into text files for analysis.

[5] I took additional notes either on my laptop or in my notebook during meetings and interviews on top of the recordings to mark my

accessed code snippets, OpenAI Playground, the Foundation's application programming interface (API) for external contractors, and the Foundation's AI applications for internal and external use.[6] I held semi-structured interviews and one-on-one chats to clarify notes and verify interpretations.

5 Case study

One of the main characters in this case is the AI reporting system (Fig. 1). Under the hood are multiple models, databases, and interfaces, all serving different purposes: The AI reporting system takes an audio-visual recording of a session, transcribes all speakers' speeches via a transcription model (TM), and then generates summaries of various formats and knowledge graphs based on the transcripts via several summarization models (SMs) and a knowledge-graph generating model (KGGM). Several iterations of the AI reporting system feature an advanced function: an AI chatbot. Users can query a chatbot via an interactive interface, asking questions about the given conference in natural languages; the chatbot generates responses that summarize session transcripts, saving users the trouble of sitting through multiple sessions. The chatbot is a retrieval-augmented generation (RAG) LLM application. First, session transcripts go through a vector-embedding model (VEM) and are stored in a vector database. Second, when the user queries the chatbot, the latter doesn't just rely on the pre-trained data of the underlying LLM; instead, it retrieves vectorized transcripts as the context that are relevant to the user's query via a retrieval model (RM). Finally, the user's initial query and the context are concatenated and sent to the chatbot's response-generating model (RGM)—its underlying LLM. This way, the chatbot answers questions using specific knowledge of the given conference.

In this case study, it is impossible to cover all facets of the AI reporting system, which has been in development for over two years at the time of writing. Instead, I select three vignettes to substantiate my claims that the developmental process and developer practices of AI models are neither mechanically objective nor devoid of sociopolitical influences, and that the ethical impacts stemming from them can only be captured via ethnographic inquiries. From the architecture to the intended capabilities of the AI model, the personal worldviews and value systems of developers and non-developers alike are embedded in this AI reporting system via developers' social and material practices.

observations and reflections on the spot. I separated reflections, which I considered the act of digesting, interpreting, and re-presenting what was being said and done during a process, from observations, which I considered the act of faithfully documenting what occurred during said process.

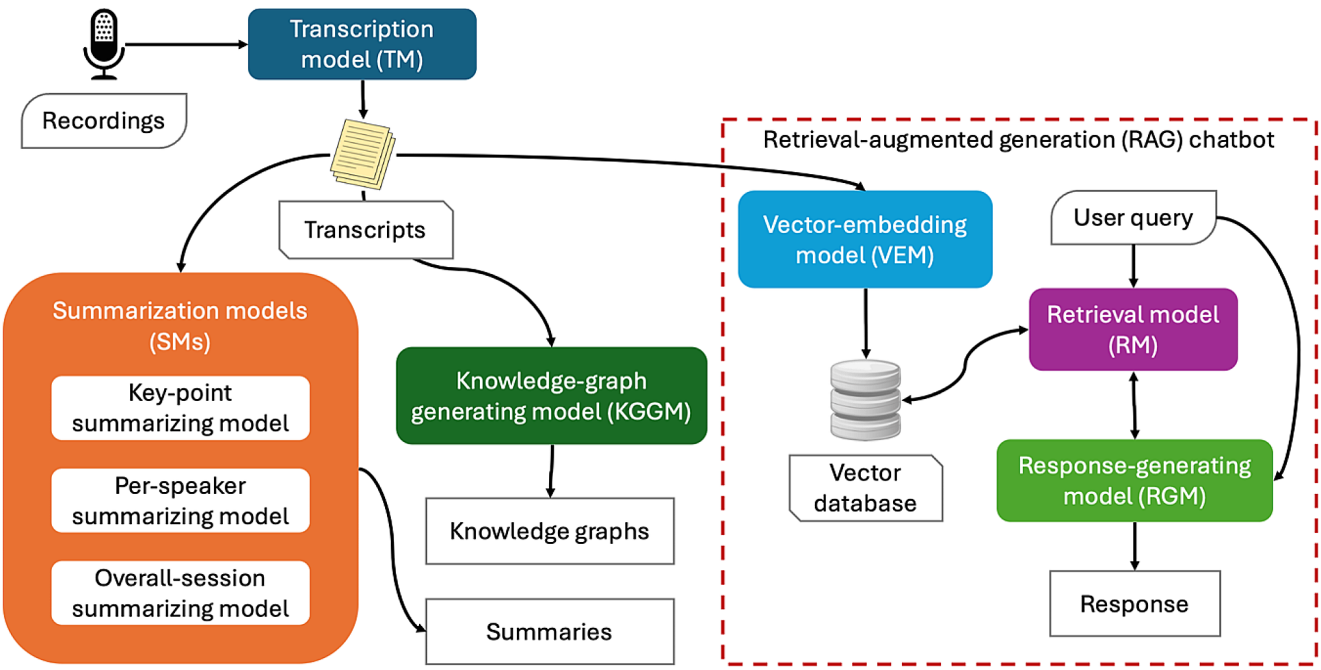

Fig. 1 Foundation's AI reporting system

[6] I could not retain copies of the technical artifacts as appendices to this paper since most are part of ongoing research and could only remain internal.

5.1 Vignette 1: the director's problem

The first vignette attempts to show that the very conceptualization of an AI model is based on the subjective imaginations of whoever is behind it. This vignette starts from the beginning, even before the technical artifacts were woven into a complex system: the Foundation's Executive Director had long held a vision for an intelligent system well before the AI lab developers began assembling one.

Back in 1992, the Director wrote his master's thesis on developing a rule-based AI system to assign legal responsibilities during international environmental accidents. In codifying international laws in an AI model, he became absorbed in the challenges different epistemes bring about and various methods to clarify fuzzy logic in a rule-based system. His interest in solving those challenges culminated in his vision of creating a knowledge management system that assists the human thinking process. He argued in *Knowledge and Diplomacy* that *knowledge management* could improve efficiency in a diplomat's work by gaining access to information, introducing a workflow that passes down information, automating routine activities during this workflow, and eventually retaining knowledge generated throughout this process [51]. At that time, he had a vision of this knowledge management system but not the means to build it.

Fast-forward to 2020, the Director had long established the Foundation, whose work areas included reporting from international conferences on digital policymaking and diplomacy. For many years, the Foundation employed human reporters to write just-in-time session summaries; however, the Director noticed an opportunity for creating a knowledge management system that could efficiently process a massive amount of information generated from such events and turn it into knowledge. Under his direction, the Foundation's AI researchers began experimenting with using AI models to summarize events and streamline the reporting process.

Recall that the "theoretical framework or the creators' interpretation of the task, problem, or issue the system is designed to address" will "inform key dimensions such as model architecture, data selection and processing, as well as the outputs." [38]. In the Director's envisioned knowledge management system, the technology product must not be a mere standalone tool but a part of a *workflow that processes information*.[7] The AI lab's task, as per instructions, was not to create a transcribing tool or a summarizing tool; the lab was instructed to create a system that *simulated and automated the entire workflow of reporting*: the system *accesses information* in the form of audio-visual recordings, *passes* such recordings to the TM that outputs session transcripts, and then *passes* such transcripts to several SMs that output session summaries in various formats like talking points or per-speaker summaries. The *workflow* then splits into multiple streams, as the summaries could be directly sent to the Foundation's or partner organizations' websites or passed down to the KGGM or the chatbot. The Director's vision of a knowledge management system elevated interoperability among multiple components to be a key dimension in the design; the inputs and outputs of each underlying model must conform to the same format for the seamless operation of the overall system.

Second, this particular way of problematizing reporting—*automation of procedures through workflow*—implicitly requires that *reporting activities be routinized*. According to the Director, a knowledge management system can automate activities as long as the latter can be logically described [51]. Therefore, the AI lab must describe "what reporting is" in serialized steps. As shown in Sect. 5.2., the AI lab resorted to finding abstract characteristics of human conversations and attempting to design generalizable rules that machines could follow when summarizing sessions. The *routinization of reporting activities* limits reporting to only the general steps that can be serialized and automated; missing from these steps are some spontaneous activities a human reporter might've taken, including researching online for additional information, emailing panelists for clarifications, or consulting colleagues for expertise.

From the first vignette, I show that conceiving an AI model that performs anything is a personal endeavor with subjective interpretations. The Director's vision of creating a knowledge management system dictated, on a higher level, the task that the AI lab needed to tackle and further defined what the resulting AI reporting system was and did. AI models are not merely machines that solve our problems; they are our imagination of the world and its problems; they are our respective worldviews, personalities, ambitions, and desires reformulated, reconfigured, and translated into codes.

5.2 Vignette 2: everyone else's problems

If the conceptualization of an AI model is personal, then one must ask who else is involved in such conceptualization and how different worldviews interact in creating the resulting model. In the second vignette, I show that the resulting AI reporting system was further refined and negotiated by a network of actors and their respective worldviews.

Just like any AI lab in an institution or a company, the Foundation's AI lab does not exist in a vacuum; instead, the

[7] From here onwards to the end of Sect. 5.1., the phrases in italics are those rephrased or borrowed from the Director's book *Knowledge and Diplomacy*, mainly from pages 8–9 where he describes a knowledge management system.

AI researchers work very closely with other teams performing different organizational functions. The AI lab is based in the Foundation's Belgrade office, sharing spaces with four other teams: The course team prepares various courses that the Foundation delivers to diplomats or higher-education students. The reporting team takes care of the daily monitoring and updating of digital news and global policy trends on the Foundation's website; it is also the team that used to provide live coverage of major international events, such as the UN General Assembly (UNGA) and Internet Governance Forum (IGF). The creative lab designs social media campaigns and visuals for all published material. The tech team manages the technical infrastructure, from websites to internal tools and applications, that allows the Foundation to function.

While the AI lab focuses on the research and design (R&D) of AI models and applications, the problems to which their research is supposed to provide answers often require more than writing a few lines of code to solve a mathematical puzzle. Instead, the AI lab reaches out to different actors within and beyond the Foundation to resolve those problems.

Recall the problem statement set by the Director: He wanted to create a knowledge management system that processes information and knowledge as generated from international events. This was highly abstract and open to further interpretation. The first layer of interpretation already happened when the Director and the AI lab decided on a workflow-simulating AI reporting system consisting of various models, all doing simpler tasks while producing interoperable inputs and outputs. But the problem must be further boiled down.

To routinize reporting activities, the AI lab must describe them in generalizable rules and instructions for each AI model. For example, to build an SM, the researchers must determine what the model could pick up from session transcripts (i.e., what to summarize). The AI lab manager (the Manager) conceived of these sessions as consisting of various conversations among multiple speakers; the issue was then to understand what could happen in a conversation. The AI lab brought in a linguist, who broke down conversations into *questions and answers*; the linguist taught the AI lab various question types one could pose in a conversation and ways to detect which type they were (rhetorical, open, etc.). The next step was to understand the answers; the AI lab consulted a debater who framed responses in terms of arguments, which were then understood as key points with corresponding supporting facts. Taking in these lessons, the AI lab instructed the reporting system to take each speaker's paragraphs from a transcript, extract key points and supporting evidence, and present a session summary in this format.

This layer of interpretation operationalized the abstract problem statement for a specific use case. The *knowledge management system that was supposed to simulate a workflow from accessing information to accruing knowledge* became one that *extracted key points and supporting facts from transcripts based on speeches delivered during a session*. How this interpretation process happened is crucial. Implicitly, the definitions of a *session* and the act of summarization in reporting activities were shaped by two actors' opinions: a session is understood in question-answer pairs, and to summarize is to detect the question and dissect the response into key points and facts. The people the AI lab consulted actively shaped the latter's understanding of what information was valuable to access and the particular way to create knowledge from it.

There were still other layers to the interpretation of the problem statement: what counts as *a good way to create knowledge*? In other words, what is *a good summary*? In preparation for deploying the AI reporting system for IGF 2023, the AI lab conversed with their colleague, the lead reporter (the Reporter) from the reporting team to learn what she would need from the AI-generated reports. It turned out that, to her, the most valuable information would not be *what was said this year* but *what was said this year that was different from all previous years*. The Reporter had years of experience covering IGF events; she already possessed knowledge of past main discussion points. Knowing that the main messages of such events usually vary little from year to year, she found only novelties in ideas or arguments valuable to her; such information would help her write a final report that identifies emerging trends in digital policy discussions. The way that the AI reporting system should transform information into knowledge now includes *comparing current knowledge to historical knowledge*, and this became a part of the AI lab's ongoing research, especially as the Foundation launched the IGF Knowledge project aiming to do just that after the end of IGF 2023.

Another person who affected the evaluation of good summaries was me. During my inquiry, I participated in the AI lab's research activities, finding ways to allow the SM to recognize more contexts and inter-relations between speaker's speeches in a transcript. I was conscious that how I approached the task was only informed and shaped by my experiences of attending sessions at a handful of conferences and events. I generalized my learnings about the usual flow of conversation in a moderator-led panelist discussion, drew a few flowcharts (Figs. 2 (left) and (right)), and then coded a script instructing an SM to extract information accordingly.

Whether my and the Reporter's beliefs about what counts as a good summary were widely shared is beside the point; likewise, whether the linguist's and debater's anatomy of a conversation was accurate does not matter. What matters

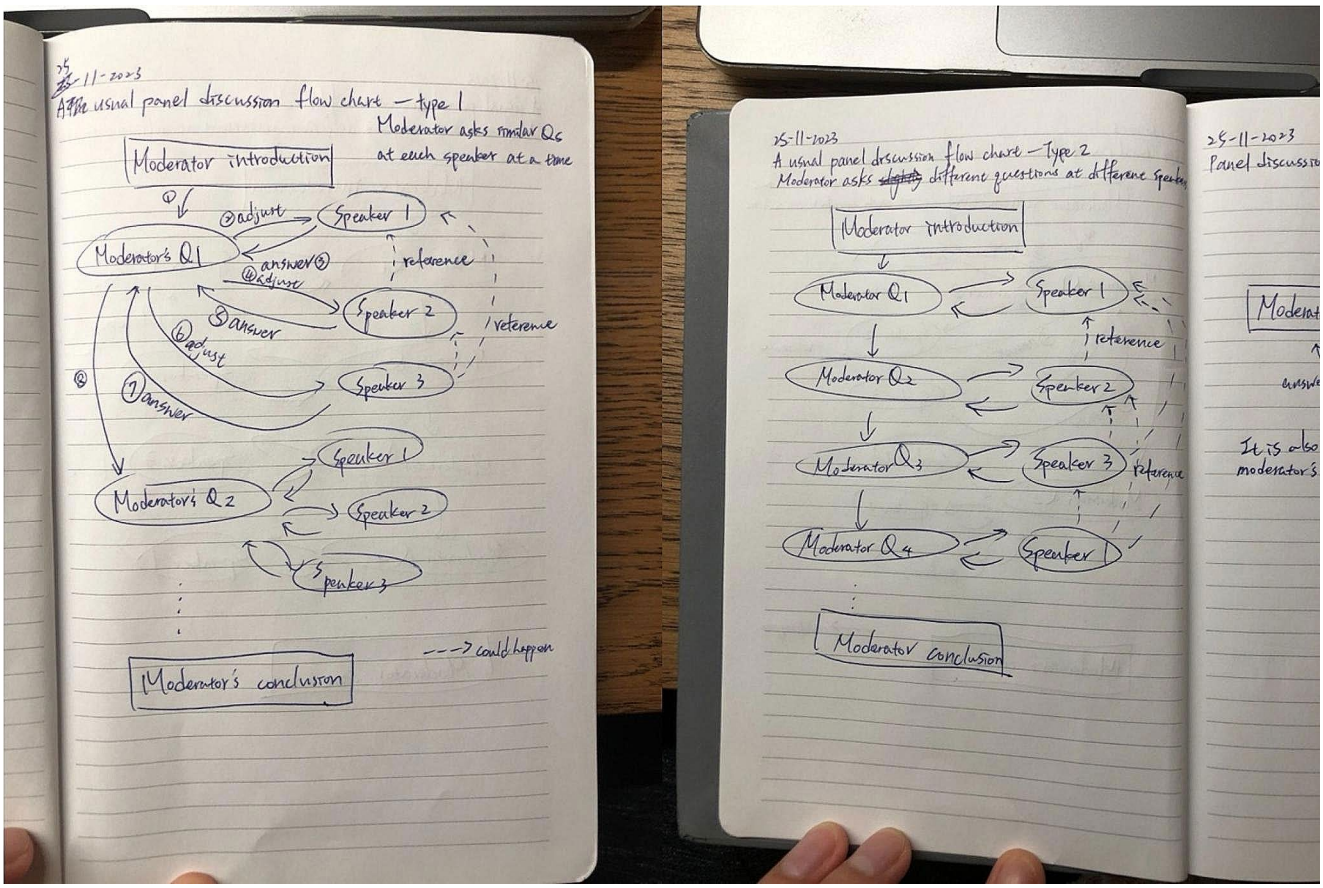

Fig. 2 (left) and (right) Flowcharts of a panelist discussion

here is that the initial problem statement was transformed into something specific and operationalized according to a handful of actors' comprehension of the problem, the different interpretations they adopt for what a session is, their personal experiences dictating which information is important, and so on and so forth. Then, the AI lab translated these expert insights into something they could work with, something that machines could work with. The comments of the linguist, the debater, and the Reporter deeply affected how the AI lab described *reporting* to the AI models and, therefore, affected what the AI reporting system did, does, or will do in the future.

The second vignette demonstrated the social nature of the supposed technical problem that the reporting system was tasked with solving. An AI model developed in a lab is never shielded from the broader social environment; the very developers who work on the model interact, exchange, and learn from other actors who all hold their worldviews and values and thereby mediate and embed all these varying worldviews and values into the AI model. An AI model's capacity is shaped by a network of actors and the worldviews they respectively hold, and the social relations among the actors further determine how these worldviews merge and become infused into the AI model. The constitution of an AI model is as social and personal as can be.

5.3 Vignette 3: the developers' problem

The previous vignettes showcase how developers engaged in social practices with a network of actors to interpret the *problem* the AI reporting system was supposed to resolve. In the third vignette, I zoom in on the AI developers themselves—what was their own interpretation of the given problem? In other words, apart from what everyone else wished the reporting system to do, what did the developers imagine about the system's capability?

Recall the chatbot feature of the AI reporting system that was based on the RAG technique. In a sequence I called the "RAG experiment," where the Manager and an AI lab member were building an RAG pipeline, I observed their interpretations of *good system performance* (i.e., what AI models should do) emerge as they hit a nail.

The lab researchers were well aware of the imperfections of the AI reporting system. In the organization, they had the clearest vision about what the reporting system could or could not do; their desire to perfect the system to their

expectations drove their motivation for conducting further experiments in prompting and RAG.

The chatbot feature of the AI reporting system was based on the RAG technique, which retrieved information via semantic searches (Fig. 3). Take the UNGA 78 AI Chat as an example: when a user asked a question, the AI Chat generated answers based on country statements made during UNGA 78 and provided a clear source. Behind the scenes was a mathematical transformation of texts called vectorization. All the UNGA 78 session transcripts first went through the VEM, which split the transcripts into text chunks by a desired size, such as paragraphs. Based on how semantically similar these text chunks were, the VEM assigned a directional value that showed the distance among these text chunks in a high-dimensional space. This formed the vector database. When a user asked, "What did the US say about cybersecurity?" The RM would calculate the semantic similarity between this question and the vectorized text chunks; going through the vector database, the RM retrieved the text chunks closest—most semantically similar—to the question. In this example, the RM might find text chunks containing both "the US" and "cybersecurity." Finally, the RGM would generate a response using both the returned chunks and the initial user query.

During my period of observation, the Manager already saw a problem with the RM and frequently brought up the following example to elucidate his felt urgency on the *lack-of-context* conundrum: Imagine a text chunk including a sentence like "Donald Trump says that there should be a wall on the southern border of the US." The meaning of this sentence could not be understood unless one knows the broader context—such as if Donald Trump was incumbent when he made such a statement. If he was, then the sentence probably conveyed the official stance of the US at that time; otherwise, it would be Mr. Trump's personal stance. The Manager argued that an RM model—such as the one used by the lab at that time—only conducted semantic searches and could not capture context beyond the text chunk; it might mistake official country stances, retrieve the wrong paragraphs for the chatbot, and lead the user to believe in wrongful answers.

Teaching the RM to recognize context became the Manager's ambition. Starting in December 2023, the AI lab began a new round of research crunch where they experimented with various retrieval techniques, text transformation methods, different data types, and RAG pipeline evaluation frameworks. Given that this was ongoing research at the time of writing, I could not describe further their conclusions. Nevertheless, there are a few things to highlight.

First, the Manager's concern was of great ethical importance, although he never framed it that way. In Tsamados et al.'s mapping of the ethics of algorithms, they proposed a few categories where ethical problems arose from epistemic factors [52]: The semantic-based RM might retrieve

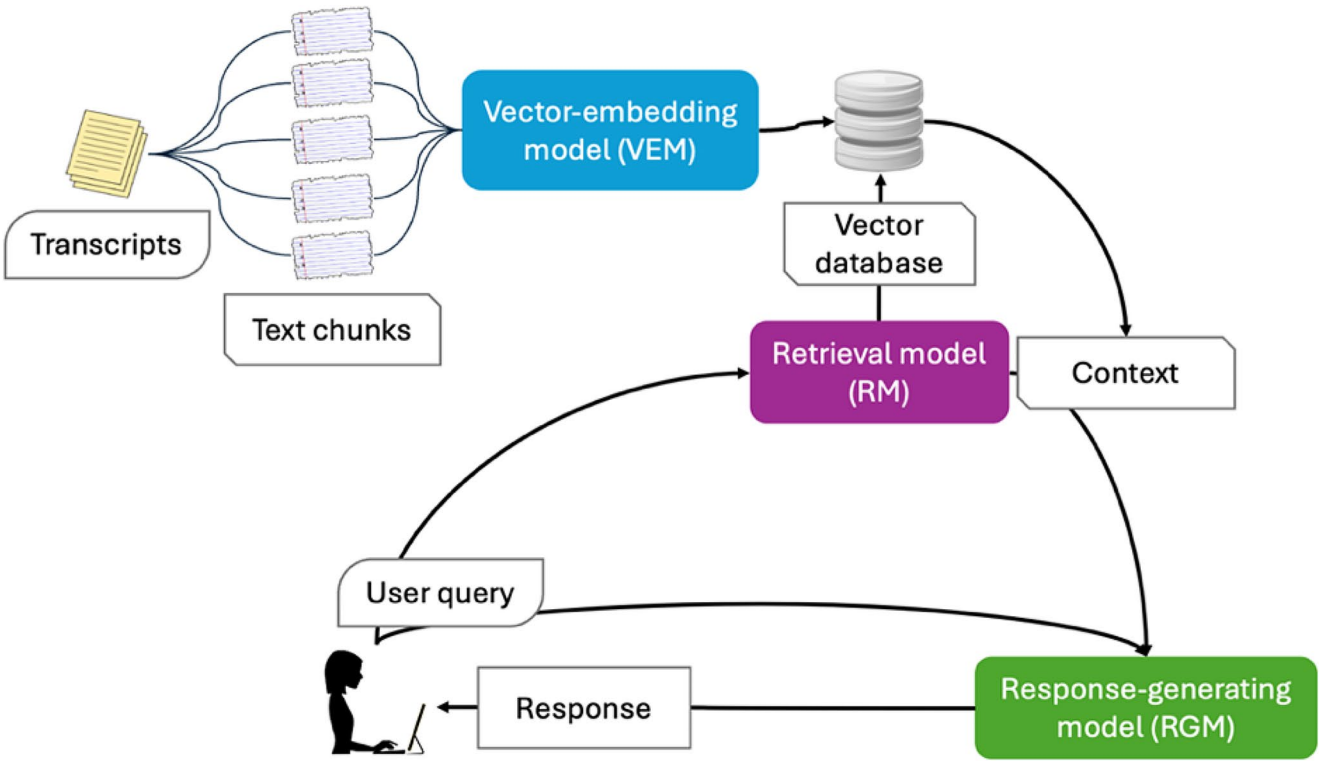

Fig. 3 RAG chatbot

paragraphs that are not relevant to the user's query because it is incapable of recognizing contexts; it might establish a wrongful connection between the search query and the vector database and commit "apophenia," the problem of "inconclusive evidence" when one sees connections where there is none [52]. This is due to the inner logic of the RM and the vectorized nature of the database—both of which decontextualize and extract the original texts from the broader context that cannot be captured in the database. To address this issue, where the chatbot hallucinates due to inconclusive evidence or invents information, the AI lab implemented a source attribution functionality: the UNGA 78 AI Chat has a "source" button that shows the retrieved paragraphs on which the chatbot generates answers. This design choice enables the users to meaningfully evaluate the quality and accuracy of the chatbot's response, thereby mitigating the potential harms of inconclusive evidence. Moreover, the source attribution functionality alleviated the problem of "inscrutable evidence"—where users could not understand how the model provided a given response—and improved the transparency of the model's operation to an extent; this further allows users to use the model's response with more trust [52]. In short, the Manager's concern and activities taken by the AI lab in response had direct ethical implications.

Second, the Manager did not frame his concern in ethical terms but instead as a *system performance* problem. In fact, most of the issues that the Manager and other AI lab members raised during meetings were described as *system performance* problems; the AI reporting system would not be considered a good system if it performed the tasks poorly or not as intended (i.e., not being able to retrieve the most accurate information in the database). Critical scholars have long documented the tendency for tech communities to understand AI models only in "rational concerns" and explain their "efficiency" and "optimality" from technical perspectives [37]. Such traditions of adopting mechanically objective views are deeply rooted in computer science and AI development. After all, the education and training of programmers are fraught with scientific papers, reports, and textbooks that follow "step-by-step maxims of conduct" or provide only "purified accounts" of scientific results [40]; all of this encourages programmers to forego "other knowledge about algorithms– such as their applications, effects, and circulation." [41].

Third, suppose the first and second points both hold; then as the Manager addressed his concern framed in system performance terms, he was actually addressing ethical concerns about the AI reporting system. This leads me to a provocative argument: one cannot meaningfully separate the technical and ethical concerns; in reality, every design decision about an AI model is both technical and ethical. During model development, AI developers naturally encounter situations where their considerations about a design choice raise ethical implications. These situations emerge as developers deliberate about their options, take an erroneous turn, calibrate their course, and eventually lead down a particular path while leaving other routes behind. In this sense, ethics is *practiced* as it is contended, intentional or not, during the developers' daily operations.

Consolidating what I have shown so far in the three vignettes, I argue that ethnographic inquiries akin to AI lab studies can reveal one critical source of AI models' ethical impacts on the world: Developer practices. In vignettes 1 and 2, I showed that the very conceptualization of an AI model and the decisionmaking process around its development are highly subjective endeavors; in deciding what the AI model will be built to do, a network of actors, including both developers and non-developers, negotiate and compromise about their desires, needs, preferences, values, and worldviews. Actors with varying levels of social status also enjoy different degrees of influence over the model, with the Director being able to set high-level objectives and architecture of the AI reporting system and the others only making suggestions about the specifics of underlying models. With these power differentials, the developers' practices in consulting others to interpret what the AI reporting system is supposed to do are not only social but political. In vignette 3, I further demonstrated that each of the developers' technical design decisions is actually ethical; the material practices they engage in can be examined as ethical. Coupled with the fact that developers' decisionmaking process is often influenced by the social network of which they are a part, the personal, social, and even political nature of developer practices revolved around building an AI model becomes evident.

6 Proposal: ethnographic audit trails

If the case study successfully proves the presence of various sociopolitical forces behind the constitution of AI models, then policymakers and ethicists face a challenge. Current *top-down* and *post-hoc* methods of EBA cannot capture developer practices in such great detail to expose the sources of the ethical implications of AI models. A *bottom-up* and *in-progress* approach to observing the ethical practices of developers—understood as every social and material practice related to the constitution of AI models—must be introduced as a potential EBA method. Although the methodologies of AI lab studies seem sufficient for the task, they remain too flexible to be standardized in regulations without more solid frameworks.

Fortunately, I can borrow a category of practices already existent in many programming projects: keeping an audit trail. Not to be confused with the auditing practices described in Sect. 2., an audit trail is a log of all steps taken to develop a specific system [13]. It is commonly used in the design of safety-critical systems such as commercial aircraft and financial industries, where step-by-step records of all decisions taken and resulting outcomes are kept [13, 53]. In programming projects, the concept of audit trails is embodied by version control tools like GitHub and GitLab, which allow programmers to establish the traceability of changes made to all individual documents [13].

There are also precedents of keeping audit trails in the AI industry. Meta AI kept a chronical log as it trained its Open Pre-Trained (OPT) model [54]; Microsoft included an audit trail for their Azure AI Health Bot [55]. Some professional services firms and AI auditing companies are also offering automated audit trail tools [20, 56]. Scholars have proposed to develop and standardize the requirements for AI audit trails: AI developers must provide chronological documentary evidence of the development of AI systems, including its problem definition, intended purpose, and design specifications [13, 35].

Complementary to the existing practices, I propose what can be called "ethnographic audit trails (EATs)" to be performed not by developers but by social studies researchers, ethicists, or anthropologists embedded in AI labs. The existing practice of AI audit trails, especially the logs automatically kept by software, focuses primarily on *what changes were made to the technical artifacts like coding scripts or datasets* instead of *the wider contexts in which such changes happen*—meaning, the sociopolitical forces described in my case study are left out. An ethnographer should keep a distinct audit trail that contextualizes the decisionmaking process and critically examines the choices taken in an AI lab in relation to the constitution of a model from a sociological and ethical point of view. A typical AI lab study captures more than just the technical artifacts but also the social environment in which such artifacts are brought into being; likewise, an EAT is essentially a lab study structured around the chronological development of AI models that enriches the content of a regular AI audit trail.

The benefits of EATs are at least three-fold: First and foremost, EATs supply the empirical data needed to complement EBA methods and enable a complete life-cycle evaluation of AI models. Since an ethnographer undertaking an EAT starts with a bottom-up position—not assuming that a particular preset of ethical values applies in a given case, they will naturally generate a rich amount of empirical data throughout their inquiry. Such information produces great contexts for understanding how exactly ethical principles are considered on the ground and how high-level ethical principles can be interpreted on a granular level. Furthermore, the empirical data of EATs can serve as evidence for post-hoc auditing activities or any other governance compliance mechanisms under EBA. In this view, the approach of EATs does not contradict but instead complements existing EBA methods.

Second, the rich amount of contextual data generated by EATs can update the interpretation and operationalization of abstract ethical principles or even add new principles. EATs can capture unknown ethical and sociopolitical risk scenarios during AI development. By documenting the entire developmental process of AI models, an ethnographer might be able to identify emerging risks that are not foreseen given the current lack of documentation of developer practices in AI labs, thereby revealing the need to develop a new set of ethical principles. Moreover, the rapid pace of AI development calls for frequent updates to existing principles, such as what they might mean and how they can be applied in new applications or frontier developments. As innovations occur in AI labs, EATs allow ethicists to follow up with the moving boundary of AI development.

Lastly, given the position of an ethnographer, EATs hold the potential to promote a culture of ethical deliberations. Ethnographers carrying out EATs are embedded in a lab setting where they join meetings, participate in tasks, and exchange personal views with the people in and around the AI lab; these ethnographers can become integrated into the specific workplace culture, approach AI model development with interpersonal perspectives, and bring in more value-sensitive thinking into conversations. Recall vignette 3 (Sect. 5.3.): If every technical decision is simultaneously ethical, and if AI developers are essentially exercising ethics in their mundane yet daily operations, then the challenge of ensuring ethical designs of AI models must be resolved within the AI lab—where AI models are designed and developed. However, this challenge may be more difficult if developers never understand their AI models as ethically non-neutral and their practices as unobjective and sociopolitically significant; even if they do, they might still lack the vocabulary to frame their concerns in ethical terms and address emerging risks. By having ethnographers participate directly in AI lab activities and even presenting their findings and reflections periodically to the very actors they observe, ethnographers are best suited to inculcate AI labs with a culture of ethical deliberations that encourages developers to always consider the potential sociopolitical consequences of their actions and make decisions under the guidance of ethical principles. The notion of fostering a culture of ethics in AI development has gained traction [57], and it certainly aligns with various schools of design methods. Notable schools like human-centered design (HCD) and value-sensitive design (VSD) aim to sensitize tech

developers to human and social concerns beyond mechanical functionalities [58–60]; with EATs, developers are further prompted to reflect in this regard, which may increase their sensibility about particular HCD methods such as inclusive and participatory design practices where different user groups and conventionally excluded communities are invited into design and decisionmaking processes [58, 61–63]. In other words, by promoting a culture of ethical discussions among developers, EATs may motivate other well-established ethical design methods.

The proposal of EATs does not overthrow the need for existing EBA methods; in reality, EATs can complement EBA and substantiate high-level abstract principles. It is also recognized that EATs may be more resource-intensive and less scalable in comparison to other EBA methods, such as metricizing abstract values to create easily applicable benchmarks across models. However, the value of EATs does not lie in their scalability or capability of offering immediately comparable insights across cases; instead, EATs are most valuable when applied to domains where AI model development is nascent, fast-moving, or high-staked, such as AI for health, environment, humanitarian aid, or legal systems.

7 Conclusions

The sociopolitical impacts of AI models are indisputable, and their growing applications in different domains give rise to a sense of urgency for us to observe, identify, and mitigate such impacts. Researchers, policymakers, and ethicists have called for using high-level abstract value-based principles to guide the development of AI models. EBA arises as a natural instrument with which we can examine whether abstract principles are respected during AI models; however, I argue that existing EBA methods take a *top-down* and *post-hoc* position vis-à-vis AI development, which are not sufficient in capturing the developer's social and material practices that encode the former's personal values and worldviews into AI models in the first place. To account for the very developmental process of AI models and capture how the sociopolitical forces around the model are embedded in it, I propose adopting the methods of EAT to generate empirical data of such process.

Moving forward, there are two challenges to be tackled: The first is to testify and further substantiate the methodologies of an EAT. The current paper uses three vignettes in a case study to necessitate a modified method of empirical observation; however, the EAT has yet to be tested in the AI development scene. There remain questions about its feasibility and adaptability in all sorts of cases. A computer-automated audit trail in programming projects can easily keep track of all minor changes down to each line of code; an ethnographer-kept audit trail may still not be able to account for the wider contexts in which each minor change is made. An ethnographer carrying out such a task must acquire sufficient experience working with AI developers, programming languages, and many other technical artifacts to gain an intuition about which technical changes and design choices are significant, for which expansive coverage is needed, and which others are less significant. In other words, how exactly an ethnographer can carry out EATs and identify foci so as to facilitate proper documentation remains to be tested.

The second is to gain access to AI labs. Given the proprietary and lucrative nature of AI models and the fast-paced development of the field, AI labs may not always welcome an outsider to participate in their daily operations. One of the reasons I could gain access to the Foundation's AI lab was my minimal capability of coding, which allowed me to become a useful member of the lab—or an *insider*. This status granted me the right to interact with some technical artifacts typically untouchable by non-programmers and non-members. Moreover, judging by the extent to which I have access, I detected there to be a "skill-to-access" ratio: the more programming skills one possesses to participate in difficult lab tasks, the more trust and membership privileges one is given, and the more access to technical artifacts one can be granted. For an ethnographer to carry out an EAT and partake in AI labs' daily operations, one might need to overcome thresholds according to the skill-to-access ratio. The closer one wishes to be to the source of AI models' sociopolitical power, the more one must understand about AI technologies.

To conclude, this paper identifies a niche yet to be filled in the wider debate about how to account for AI models' sociopolitical impacts. The bottom-up and in-progress position that an EAT offers may complement existing EBA methods and ensure ethical considerations throughout the life cycle of AI model development.

Acknowledgements The author wishes to thank Florian Jaton for helpful comments on earlier versions of this article. The author also wishes to thank Jérôme Duberry and Oana Ichim for their guidance during the master's thesis research. Finally, the author wishes to extend gratitude towards the organization where the inquiry takes place.

Funding Open access funding provided by Geneva Graduate Institute.

Declarations

Consent for publication Written informed consent for publication of the personal details and/or images and/or audio recordings and/or videos was obtained from the participant/parent/guardian/relative of the participant. A copy of the consent form is available for review by the Editor of this journal.

Conflict of interest The author did not receive any form of financial support from any organization for the submitted work. The author was doing a paid internship with the organization featured in the case study while carrying out research for this submitted work. However, the research was academically independent, as it was conducted as part of the author's thesis research at the Graduate Institute of International and Development Studies instead of in his capacity as an intern at the case study organization.

References

1. Suchman, L., Gerst, D., Krämer, H.: 'If you want to Understand the Big issues, you need to Understand the Everyday practices that constitute them.' Lucy Suchman in Conversation with Dominik Gerst & Hannes Krämer. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research. (2019). https://doi.org/10.17169/fqs-20.2.3252
2. Panch, T., Mattie, H., Celi, L.A.: The 'Inconvenient truth' about AI in Healthcare. Npj Digit. Med. **2**(1), 1–3 (2019). https://doi.org/10.1038/s41746-019-0155-4
3. Richardson, R., Schultz, J., Crawford, K.: L Rev. Online. **94**, 192–228 (2019). https://ssrn.com/abstract=3333423 Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. 94 N.Y.U.
4. Yong, E.: A Popular Algorithm Is No Better at Predicting Crimes Than Random People. The Atlantic (blog). (2018). https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/ Accessed 20 June 2024
5. Dressel, J., Farid, H.: The Accuracy, Fairness, and limits of Predicting Recidivism. Sci. Adv. **4**(1) (2018). https://doi.org/10.1126/sciadv.aao5580
6. Ferrer, X., van Nuenen, T., Such, J.M., Coté, M., Criado, N.: Bias and discrimination in AI: A cross-disciplinary perspective. IEEE Technol. Soc. Mag. **40**(2), 72–80 (2021). https://doi.org/10.1109/MTS.2021.3056293
7. Bartoletti, I.: An Artificial Revolution: on Power, Politics and AI. Indigo, London (2020)
8. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven, Connecticut, US (2022)
9. Dixon, R.B.L.: A principled governance for emerging AI regimes: Lessons from China, the European Union, and the United States. AI Ethics. **3**(3), 793–810 (2023). https://doi.org/10.1007/s43681-022-00205-0
10. Mäntymäki, M., Minkkinen, M., Birkstedt, T., Viljanen, M.: Defining organizational AI Governance. AI Ethics. **2**(4), 603–609 (2022). https://doi.org/10.1007/s43681-022-00143-x
11. Radu, R.: Steering the governance of Artificial Intelligence: National Strategies in Perspective. Policy Soc. **40**(2), 178–193 (2021). https://doi.org/10.1080/14494035.2021.1929728
12. Stix, C.: Actionable principles for Artificial Intelligence Policy: Three pathways. Sci Eng. Ethics. **27**(1), 15 (2021). https://doi.org/10.1007/s11948-020-00277-3
13. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., et al.: Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv. (2020). https://doi.org/10.48550/arXiv.2004.07213
14. Mökander, J., Morley, J., Taddeo, M., Floridi, L.: Ethics-based auditing of automated decision-making systems: Nature, Scope, and limitations. Sci Eng. Ethics. **27**(4), 44 (2021). https://doi.org/10.1007/s11948-021-00319-4
15. Mökander, J., Floridi, L.: Ethics-based auditing to develop trustworthy AI. Mind. Mach. **31**(2), 323–327 (2021). https://doi.org/10.1007/s11023-021-09557-8
16. Brown, S., Davidovic, J., Hasan, A.: The Algorithm audit: Scoring the algorithms that score us. Big Data Soc. **8**(1) (2021). https://doi.org/10.1177/2053951720983865
17. Ayling, J., Chapman, A.: Putting AI Ethics to work: Are the Tools fit for purpose? AI Ethics. **2**(3), 405–429 (2022). https://doi.org/10.1007/s43681-021-00084-x
18. Information Commissioner's Office: Guidance on AI and Data Protection. ICO. (2023). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ Accessed 20 June 2024
19. PricewaterhouseCoopers (PwC): Responsible AI Toolkit. PwC. (2024). https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html Accessed 7 April 2024
20. Ernst: & Young (EY).: Responsible AI. EY. https://www.ey.com/en_ch/ai/responsible-ai. Accessed 7 April 2024
21. Deloitte: Trustworthy Artificial Intelligence (AI)™. Deloitte United States. https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html. Accessed 7 April 2024
22. Holistic, A.I.: What Is AI Auditing? Holistic AI. (2022). https://www.holisticai.com/blog/ai-auditing Accessed 7 April 2024
23. Fiddler, A.I., Observability, A.I., Model Monitoring, M.L., LLMOps, and, Explainable, A.I.: April. Fiddler AI. (2024). https://www.fiddler.ai/. Accessed 7
24. Arthur: Observability. Arthur. (2024). https://www.arthur.ai/solution/observability. Accessed 7
25. Parity Consulting: Parity Consulting. Parity Consulting. https://www.get-parity.com. Accessed 7 April 2024
26. Pessach, D., Shmueli, E.: A review on Fairness in Machine Learning. ACM Comput. Surveys. **55**(3), 1–44 (2022). https://doi.org/10.1145/3494672
27. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. arXiv. (2022). https://doi.org/10.48550/arXiv.1908.09635
28. Mökander, J., Floridi, L.: Operationalising AI governance through Ethics-based auditing: An industry case study. AI Ethics. **3**(2), 451–468 (2023). https://doi.org/10.1007/s43681-022-00171-7
29. AstraZeneca.: Advancing Data and Artificial Intelligence. AstraZeneca. (2020). https://www.astrazeneca.com/sustainability/ethics-and-transparency/data-and-ai-ethics.html Accessed 20 June 2024
30. John-Mathews, J.-M., Cardon, D., Balagué, C.: From reality to World. A critical perspective on AI Fairness. J. Bus. Ethics. **178**(4), 945–959 (2022). https://doi.org/10.1007/s10551-022-05055-8
31. Lee, M.S.A., Floridi, L., Singh, J.: Formalising Trade-Offs beyond Algorithmic Fairness: Lessons from ethical Philosophy and Welfare Economics. AI Ethics. **1**(4), 529–544 (2021). https://doi.org/10.1007/s43681-021-00067-y
32. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model Cards for Model Reporting. In Proceedings of the Conference on Fairness,

Accountability, and Transparency. 220–29 (2019). https://doi.org/10.1145/3287560.3287596
33. OpenAI: GPT-4V(Ision) System Card. (2023). https://openai.com/research/gpt-4v-system-card Accessed 20 June 2024
34. MetaAI.: System Cards, a New Resource for Understanding How AI Systems Work. (2023). https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/. Accessed 22
35. Mökander, J., Schuett, J., Kirk, H.R., Floridi, L.: Auditing large Language models: A Three-Layered Approach. AI Ethics. (2023). https://doi.org/10.1007/s43681-023-00289-2
36. Marino, M.: Critical Code Studies. The MIT Press, Cambridge, MA, US (2020)
37. Kitchin, R.: Thinking critically about and researching algorithms. Inform. Communication Soc. **20**(1), 14–29 (2017). https://doi.org/10.1080/1369118X.2016.1154087
38. Ugwudike, P.: AI audits for assessing design logics and building ethical systems: The case of predictive policing algorithms. AI Ethics. **2**(1), 199–208 (2022). https://doi.org/10.1007/s43681-021-00117-5
39. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P.: Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, 33–44 (2020). https://doi.org/10.1145/3351095.3372873
40. Jaton, F.: The Constitution of Algorithms. The MIT Press, Cambridge, MA, US (2021)
41. Seaver, N.: Knowing algorithms. In: Vertesi, J., Ribes, D. (eds.) digitalSTS: A Field Guide for Science and Technology Studies, pp. 412–422. Princeton University Press, Princeton, New Jersey, US (2019)
42. Diakopoulos, N.: Transparency. In: Dubber, M.D., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI, pp. 197–213. Oxford University Press, Oxford, UK (2020)
43. Seaver, N.: Algorithms as culture: Some tactics for the ethnography of Algorithmic systems. Big Data Soc. **4**(2) (2017). https://doi.org/10.1177/2053951717738104
44. Montfort, N., Baudoin, P., Bell, J., Bogost, I., Douglass, J.: 10 PRINT CHR$(205.5+RND(1));: GOTO 10. The MIT Press, Cambridge, MA, US (2012)
45. S Geiger, R.: Bots, Bespoke, Code and the Materiality of Software platforms. Inform. Communication Soc. **17**(3), 342–356 (2014). https://doi.org/10.1080/1369118X.2013.873069
46. Latour, B., Woolgar, S.: Laboratory Life: The Social Construction of Scientific Facts. Sage, Beverly Hills, LA, US (1979)
47. Knorr Cetina, K.D.: Laboratory studies: The Cultural Approach to the study of Science. In: Jasanoff, S. (ed.) Handbook of Science and Technology Studies, pp. 140–167. Sage, LA, US (1995)
48. Knorr Cetina, K.D.: The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science. Pergamon, Oxford (1981)
49. Lynch, M.: Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory. Routledge Kegan & Paul, Boston, US (1985)
50. Traweek, S.: Beamtimes and Lifetimes: The World of High Energy Physicists. Harvard University Press, Cambridge, MA, US (1988)
51. Kurbalija, J.: Knowledge Management and Diplomacy. In: Kurbalija, J. (ed.) Knowledge and Diplomacy, pp. 7–19. Academic Training Institute, Msida, Malta (2002)
52. Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., Floridi, L.: The Ethics of algorithms: Key problems and solutions. AI Soc. **37**(1), 215–230 (2022). https://doi.org/10.1007/s00146-021-01154-8
53. Investopedia: Audit Trail: Definition, How It Works, Types, and Example. Investopedia. (2024). https://www.investopedia.com/terms/a/audittrail.asp. Accessed May 26
54. Meta Research: Metaseq/Projects/OPT/Chronicles at Main. Facebookresearch/Metaseq. (2024). https://github.com/facebookresearch/metaseq/tree/main/projects/OPT/chronicles. Accessed May 26
55. Microsoft Learn: Audit Trails - Azure AI Health Bot. (2023). https://learn.microsoft.com/en-us/azure/health-bot/audit-trails Accessed 20 June 2024
56. Credo, A.I.: Credo AI Audit Trail. Glossary. (2024). https://www.credo.ai/glossary/credo-ai-audit-trail. Accessed May 26
57. Langlois, L., Dilhac, M.-A., Dratwa, J., Ménissier, T., Ganascia, J.-G., Weinstock, D., Bégin, L., Marchildon, A.: L'éthique au cœur de l'IA., Obvia: Quebec, Canada (2023). https://www.obvia.ca/ressources/lethique-au-coeur-de-lia
58. Auernhammer, J.: Human-Centered, A.I.: The role of human-centered Design Research in the development of AI. DRS Bienn. Conf. Ser. (2020). https://doi.org/10.21606/drs.2020.282
59. Sadek, M., Calvo, R.A., Mougenot, C.: Designing Value-Sensitive AI: A critical review and recommendations for Socio-Technical Design processes. AI Ethics. (2023). https://doi.org/10.1007/s43681-023-00373-7
60. Friedman, B., Hendry, D.G.: Value Sensitive Design: Shaping Technology with Moral Imagination. The MIT Press, Cambridge, MA, US (2019)
61. Bjerknes, G., Bratteteig, T.: User participation and democracy: A discussion of Scandinavian Research on System Development. Scandinavian J. Inform. Syst. **7**(1), 73–98 (1995). https://aisel.aisnet.org/sjis/vol7/iss1/1
62. Neuhauser, L., Kreps, G.L., Morrison, K., Athanasoulis, M., Kirienko, N., Van Brunt, D.: Using design science and ariicial intelligence to improve health communicaion: ChronologyMD case example. Paient Educaion Couns. **92**(2), 211–217 (2013). https://doi.org/10.1016/j.pec.2013.04.006
63. Abascal, J., Nicolle, C.: Moving towards inclusive design guidelines for socially and ethically aware HCI. Interacing Computers. **17**(5), 484–505 (2005). https://doi.org/10.1016/j.intcom.2005.03.002