

Context and Symmetry in Auditing: A Case Study of Skeleton Inference in Motion Capture

Emma Harvey^{*1}, Emanuel Moss^{*2, 3}, Hauke Sandhaus¹, Abigail Z. Jacobs⁴, Mona Sloane²

¹Cornell Tech

²University of Virginia

³Intel Corporation

⁴University of Michigan

evh29@cornell.edu, emanuel.moss@intel.com, hgs52@cornell.edu, azjacobs@umich.edu, mona.sloane@virginia.edu

Abstract

Humans are increasingly expected to interact with AI systems that observe and make inferences about them—but do these systems actually work? A standard approach to answering this question is AI auditing. Conducting an AI audit requires identifying how a system behaves (i.e., determining what types of inputs to audit it with and then observing and documenting actual system behavior) and contrasting that with how a system should behave (i.e., determining what the nominal outputs of a system should look like). We argue that this is best done through a **contextual audit**, which we introduce as a method for auditing measurements within the context of the practices that produce them. We show how contextual auditing enables the interrogation of assumptions implicit in the audit process and allows auditors to be explicit about what serves as *ground truth*, which we define as verifiable measurements about the real world against which systems are evaluated. We outline how the concept of **symmetry** from science and technology studies can enable audits when ground truth is unknown, unknowable, or contested. Finally, to demonstrate how contextual and symmetric audits can be conducted in practice, we present a case study of skeleton inference in motion capture and propose areas suggested by our case study as particularly fruitful for future audits of motion capture systems.

actual outputs of a system to *nominal* expected outputs—in other words, comparing how the system does behave to how it should behave (Birhane et al. 2024; Mei et al. 2026; Sandvig et al. 2014; Sloane and Moss 2023). This immediately raises two questions. First, **what inputs are relevant for identifying patterns in a system’s actual outputs?** Second, **what should a system’s nominal outputs look like** (Mei et al. 2026)?

A typical approach to AI auditing is to elicit outputs from a system in a systematic way to capture a specific behavior (Bandy 2021). Researchers often use pre-existing or synthetic data as inputs to produce outputs that can be analyzed to assess how well they match nominal outputs (e.g., Fleisig et al. 2024; Hofmann et al. 2024; Imana et al. 2025; Wilson and Caliskan 2024). However, such approaches have been criticized for failing to capture real-world use, and thus the potential real-world impacts, of systems (Harvey, Kizilcec, and Koenecke 2025; Wang, Ho, and Koyejo 2025). Instead, drawing on Sloane et al. (2023), we argue that what is required is an auditing approach that uses inputs and nominal outputs that are meaningful within the contexts in which AI systems are actually used. To address this, we introduce the **contextual audit**, which makes claims about the accuracy of system outputs *within the context of the practices that produce them*. In this work, we lay out the importance of contextual audits and their role within the broader domain of AI auditing. Then, to demonstrate how contextual audits can be conducted, we present a case study of skeleton inference in motion capture (mocap). *Skeleton inference*, a basic task in mocap, involves sensing landmarks on the surface of a body and applying algorithms to infer the position and motion of skeletal elements within the body. Skeleton inference can be used to produce *body segment parameters* (BSPs), which are measurements of various properties (e.g., length) of body segments (e.g., forearm). BSPs are then used in a variety of mocap applications, such as workplace safety monitoring and gait recognition.

Like other forms of audits, contextual audits make claims about the behavior of a system by comparing its actual outputs to *nominal outputs*, which describe ideal behavior. A typical approach requires that a system meet some threshold for accuracy (overall or in terms of relative accuracy, e.g., across demographic groups). Determining accuracy

Appendix — <https://bit.ly/osfxy4jg>

1 Introduction

Humans are increasingly enrolled into interactions with AI systems that observe and make inferences about them. These systems, which can be thought of as making *measurements* about people’s bodies or behaviors, now moderate high-stakes outcomes like employment (e.g., candidate assessments (Landers and Sanchez 2022)) and physical health (e.g., workplace safety monitoring (Xsens 2024)). Despite the widespread adoption of AI systems, a crucial question is often left unanswered: do they work as intended (Raji et al. 2022)? A standard approach to answering this question relies on *AI auditing*, which typically entails comparing the

^{*}These authors contributed equally.

requires comparing system outputs to a *ground truth*, which we define as a verifiable set of measurements about the real world. To conduct a contextual audit of skeleton inference, we sought to assess the accuracy of BSPs produced by a mocap system. However, there is no single agreed-upon system of *ground truthing* (i.e., determining what should be taken as ground truth) in the context of BSPs. For example, should the length of a forearm be measured based on the length of the radius or ulna (neither of which are directly observable without dismemberment), or should it be based on the distance between landmarks on the surface of the body (which may change depending on forearm position or the presence of soft tissue)? Deciding what accuracy means is far from simple—but it is crucial for AI auditing. In our case study audit, we compare BSPs derived from mocap systems to those derived using tape measured anthropometric techniques. However, anthropometry, like mocap, is a practice of measurement that attempts to establish, but cannot directly measure, ground truth.

To address this, we propose that contextual audits dealing with the measurement of concepts where ground truth is unknown, unknowable, or contested must borrow from the science and technology studies (STS) concept of **symmetry** (Bloor 1991). Symmetry requires treating competing theories of truth with the same analytical approach—as if either might be true or false—and draws attention to the social practices of establishing their validity (Jasanoff 2019). We argue that symmetry can be adapted for AI audits: it guides auditors to treat the outputs of each measurement modality as provisionally true, and then to interrogate the implications of doing so. Because assessing implications of systems’ behaviors is the focus of AI auditing, symmetry enables a practice for assessing AI systems even in the absence of a ground truth. Practically, symmetry can be used to explore where measurements produced through different modalities are in tension and not easily reconcilable over time and then to point to where the assumptions underpinning a system should be further investigated. Overall, we make the following contributions:

1. We introduce **contextual audits** for auditing measurements in the context of the practices that produce them, and we show how they enable interrogating what serves as ground truth and identifying the consequences of using a given ground truth to audit an AI system.
2. We show how **symmetry** can enable audits when ground truth is unknown, unknowable, or contested.
3. We demonstrate the applicability of contextual auditing through a case study audit of skeleton inference in motion capture.

2 Background and Related Work

In this section, we briefly describe existing approaches to AI auditing (§2.1) and define the concepts from sociology and STS—social practice theory and symmetry—that underlie our work (§2.2). To ground our case study, we also discuss the practice of measurement in general and of skeleton inference in mocap in particular (§2.3).

2.1 AI Auditing

AI auditing is the practice of comparing how AI systems do behave to how they should behave, with the goals of identifying harms, creating accountability, or improving systems (Bandy 2021; Birhane et al. 2024; Becerra Sandoval and Jing 2025; Mei et al. 2026; Mökander 2023; Raji et al. 2020; Rhea et al. 2022b; Sloane 2021; Sloane, Moss, and Chowdhury 2022; Sloane and Moss 2023). Sandvig et al. (2014) emphasize that system behavior should be elicited through “a combination of [a system] and its data,” i.e., by querying a system with specific inputs and observing its outputs (Rhea et al. 2022b). This formulation of AI auditing raises two key questions. First, how do we determine how a system actually behaves—what inputs are relevant for identifying patterns in a system’s actual outputs (Mei et al. 2026)? Second, how do we decide how a system should behave—what should a system’s nominal outputs look like? Answering these questions requires considering:

1. The *context* in which AI systems are designed and used.
2. The *ground truth* according to which AI systems can be evaluated—ground truth is taken as nominal outputs.

Context in AI Auditing. Deciding what inputs to use to conduct an AI audit is a non-trivial question. Collecting real-world input data is time-consuming and expensive, and can raise user privacy concerns (e.g., if users share personal information as a part of real-world system use). As a result, auditors often adapt pre-existing data or create their own synthetic data (e.g., Fleisig et al. 2024; Hofmann et al. 2024; Wilson and Caliskan 2024). Then, audits are conducted by examining the outputs produced by using those inputs to prompt the AI system. However, this approach is inherently limited: it does not allow for the evaluation of AI systems in their normal context of use. Harvey, Kizilcec, and Koencke (2025) argue that audits should consider *ecological validity*, or “the extent to which experimental findings can generalize to the ‘real world’ situation that a researcher wishes to understand” (Kihlstrom 2021). This means that the inputs should be produced as they would be produced in a real-world application, and outputs should be analyzed according to their actual context of use. Wang, Ho, and Koyejo (2025) underscore the importance of contextualized inputs: they find that the performance of large language models (LLMs) on widely-used benchmarks depends on the chat history of the prompter, meaning that generic evaluations do not necessarily provide an understanding of performance and impact for real-world users. Looking at outputs of a system divorced from their context of use also provides a misleading understanding of system performance (Gosciak et al. 2026). For example, Pruss (2023) finds that judges rarely use a recidivism prediction tool, rendering audits of it meaningless; similarly, Cheng et al. (2022) find that audits of a child welfare assessment tool overstate racial bias because caseworkers’ selective adoption of the tool mitigates some biases. In this work, we propose the **contextual audit** as an approach for conducting AI audits that consider how systems are actually used in practice. In a contextual audit, inputs are produced following actual practices of system users, and outputs are evaluated in the context of their use.

Ground Truth in AI Auditing. An AI audit requires a comparison between nominal and actual behaviors (Sloane and Moss 2023). Researchers often test normative expectations, e.g., that systems do not encode “harmful discrimination” (Sandvig et al. 2014). A common operationalization of harmful discrimination is whether systems are equally accurate (i.e., produce actual outputs that are equally close to nominal ones) across demographic groups (Angwin et al. 2016; Buolamwini and Gebru 2018; Obermeyer et al. 2019; Wilson et al. 2021). For many systems, however, determining accuracy is challenging, or even impossible: ground truth is not directly observable either because it is a theoretical (and potentially contested) concept (e.g., creditworthiness) or because it simply cannot be directly accessed (e.g., certain anatomical measurements) (Jacobs and Wallach 2021). In other words, there is no agreed upon ground truth data—or nominal value—according to which a system can be evaluated. In these cases, it is not immediately clear how to conduct an AI audit. In the absence of ground truth, researchers have tested whether AI systems achieve parity (i.e., produce similar outputs across demographic groups) (Sweeney 2013; Feng and Shah 2022; Wilson et al. 2021) or reliability (i.e., produce repeatable outputs) (Rhea et al. 2022a,b). However, neither approach fully answers the question of whether AI systems—specifically, systems that make measurements about people—produce outputs that are *valid*, i.e., that meaningfully measure what they are intended to measure (Jacobs and Wallach 2021). In this work, we propose that the concept of **symmetry**, drawn from STS, can be a powerful tool for enabling audits in the absence of an agreed upon ground truth or in situations in which multiple ground truths are conflicting or contested (Bloor 1991).

2.2 Sociology and STS Concepts for AI Auditing

Since its inception, AI auditing has drawn from multiple disciplines. It was originally proposed as an adaptation of social science auditing, which uses empirical methods like field experiments to identify and quantify discrimination in societal systems (Bertrand and Mullainathan 2004; Sandvig et al. 2014; Vecchione, Levy, and Barocas 2021). As AI audits become increasingly important for ensuring AI systems can safely be deployed in society, researchers and practitioners have drawn lessons from other fields, including financial auditing and engineering or medical safety evaluations (Mökander 2023; Raji et al. 2020; Rismani et al. 2023). We demonstrate that AI auditing can further benefit from drawing more explicitly on additional approaches from the fields of sociology and STS.

Social Practice Theory. Contextual audits require an understanding of the practices by which inputs to an AI system are produced and outputs of an AI system are used. *Social practice theory* is an approach to examining how technical systems operate within social settings that focuses on the activities and behaviors that individuals and groups conduct when engaging with a technological object. Rather than focus on the object or the individual, social practice theory focuses on the practices that stabilize, change, and re-stabilize over time. Social practices are composed of

three elements: *materials* (objects), *competences* (skills and knowledge), and *meanings* (understandings and beliefs) that change and reassemble in specific contexts (Shove 2007; Shove, Pantzar, and Watson 2012). Social practice theory can provide a frame for understanding the contexts in which AI systems are designed and used, particularly for understanding how AI systems operate as novel materials being incorporated into already-existing sets of competences and meanings (Harvey et al. 2024; Sloane and Moss 2022).

Symmetry. When concepts can be measured in multiple ways (Hand 2004; Klein 1988), *concurrent validity* assessments are often conducted, in which one measurement model (the particular systematization and operationalization of the underlying concept) is evaluated against another (Jacobs and Wallach 2021). This raises a question: which model is the better approximation of the underlying concept? While one model can be treated as the basis for comparison with another model, either (or both) may be flawed. Treating one or the other as ground truth obscures the shortcomings of that model while also holding the other to an impossible—and potentially flawed—standard. Here, we draw upon the general principle of *symmetry*, which was originally developed for the social study of knowledge and widely deployed in STS (Bloor 1991; Jasanoff 2019). For STS scholars, the impulse to separate ‘true’ scientific descriptions of reality from ‘false’ ones is held at bay by applying a symmetrical lens to each. Each scientific description is analyzed using the same techniques, as if it has the same likelihood of being true or false. In STS, symmetry is deployed not to erase the distinction between truth and error (Fuller 2016), but to better understand how scientific claims are made and come to be accepted by a broader community. Similarly, we argue that AI auditing can adopt symmetry to examine different measurement modalities without needing to hold one as a ground truth against which the other is compared. Rather, symmetry can help analyze the social and technical practices through which systems come to be accepted as valid and reliable.

Symmetry is important for AI audits because it acknowledges the often-contingent nature of ground truth and exposes the assumptions needed to accept data as ground truth. This is crucial for identifying potential gaps between underlying concepts and how they are represented as ground truth data. If such gaps go undetected by an audit process, audits can fail to detect system harms. For example, toxicity benchmarks may produce dramatically different results depending on the populations and guidelines used to produce the benchmarks (Fleisig, Abebe, and Klein 2023; Kirk et al. 2024). Holding a symmetrical view in which each benchmark *could* represent ground truth for a given context allows auditors to unpack the assumptions underlying the data in constructive ways; e.g., to justify using one benchmark instead of another, to modify or improve a benchmarking practice, or to buttress an audit with additional investigative techniques. Crucially, symmetry frames its analysis within the context to which the audit tool is being applied by asking how the assumptions that have been made in the design of a system affect a specific implementation in a particular place, at a particular time.

2.3 Measurement and Skeleton Inference

Measurement is the quantification of an (often unobservable) underlying concept using observable data (Bandalos 2018). Measurement is distinct from prediction and related tasks like generation in that measurement is an attempt to approximate an existing ground truth (Mussnug 2022).

Mocap systems are AI systems that perform measurement tasks. They consist of sensors, which collect data about surfaces of bodies, and algorithms, which make inferences based on that data. Among the most basic tasks conducted by mocap systems is *skeleton inference*, which seeks to measure the position and motion of skeletal elements within a body.¹ Skeleton inference produces BSPs, which measure various properties of body segments. Skeleton inference and BSPs are the building blocks for a variety of mocap applications, ranging from low-stakes to high-stakes settings. For example, they enable the creation of lifelike human motion for video games. They are also used in biometric tasks like gait recognition, which seeks to identify individuals based on how they walk, and for health-related tasks, including workplace safety monitoring and physical therapy (Han and Lee 2013; Topham et al. 2022). As mocap applications proliferate, it becomes increasingly important to understand how these technologies perform across the wide range of human bodies that they might come into contact with.

There is good reason to believe that skeleton inference may encode discriminatory biases by performing differently well (and, in turn, producing differently accurate BSPs) for different types of bodies. Modern mocap systems and the inferences they make are largely built on data from a small number of thin male cadavers (Harvey et al. 2024). Prior work has found that inferences made by mocap systems do not generalize equally well to individuals whose bodies do not match those in the data (Durkin and Dowling 2003). Importantly, skeleton inference is challenging to “ground truth”: the nominal values of BSPs produced through skeleton inference are often unknown or contested (Bodenheimer et al. 1997). Concurrent validity assessments are often used to evaluate mocap systems that do skeleton inference—but Harvey et al. (2024) have found that such assessments often rest on strong assumptions and may erroneously treat flawed measurements as ground truth.

In this work, we conduct a case study audit of the BSPs produced by a mocap system. Beginning with a straightforward concurrent validity assessment, in which we compare BSPs produced by a mocap system to those produced through an alternative modality (anthropometric measurements taken with a tape measure), we show how context and symmetry are critical tools for auditors seeking to evaluate real-world performance of AI systems without access to ground truth data.

¹We refer to *skeletal elements* rather than *bones* to indicate that skeleton inference infers the position and motion of rigid structures within the body, but that these structures are typically simplified models of skeletons.

3 Framework for Contextual Audits

A **contextual audit** makes claims about outputs of AI systems *within the context of the practices that produce them*. This means that contextual audits must begin by selecting a specific application of a system—they cannot be expected to evaluate every set of outputs that a system is capable of producing. Then, contextual audits involve identifying the particular social practices that are used to produce outputs in the selected application. In this way, contextual audits help to answer the question, “what inputs are relevant for identifying patterns in a system’s actual outputs?” by suggesting that we follow the practices specific to our desired context of use. Contextual audits also surface relevant sets of assumptions underpinning the application of the AI system. These assumptions can point to *facets*, or features across which system outputs should be compared, as well as ideals describing how systems should behave across those facets. In this way, contextual audits help to answer the question, “what should a system’s nominal outputs look like?” (Rhea et al. 2022a).

While the standard approach to AI auditing centers quantitative analysis, contextual audits fundamentally require qualitative research. Understanding the practices that produce the outputs of AI systems requires deep engagement with the users who conduct those practices. A contextual audit should therefore include a field observation or an ethnography of the users who operate an AI system within a specific context; it may also include informational interviews with those users. The context in which users operate is also shaped by the AI system itself. In particular, historical decisions about how systems should be built and documentation describing how systems should be used are both factors that may be taken for granted by users, but that deeply affect their context of use. Thus, contextual audits should also include a review of the research, data, and design decisions underpinning an AI system, as well as the system’s documentation. Appropriate methods for this aspect of a contextual audit may include a systematic literature review (e.g., Harvey et al. 2024), a historical analysis (Becerra Sandoval and Jing 2025), or a trace ethnography (Geiger and Ribes 2011).

To support this work, we adapt the socio-technical matrix proposed by Sloane, Moss, and Chowdhury (2022) and extended by Rhea et al. (2022a) and Sloane, Jacobs, and Moss (2025). That matrix, originally designed to facilitate audits of automated decision-making systems used in hiring, prompts auditors to examine the context in which a system is intended to operate, interrogate how the system works in practice, and identify the assumptions and epistemological roots that underpin the system (and thus what is considered its ground truth). The matrix, which we have further adapted here, aids in identifying assumptions about what can and should be accepted as nominal values by helping auditors define, from a system’s many possible purposes, the specific goals that it is being used to achieve. Auditors can then isolate the sets of assumptions that underpin those goals and trace how those assumptions are operationalized in specific measurements produced by the system. The matrix, then, enables auditors to audit a complex AI system by clearly stipulating the specific purpose it is being used for, the capability drawn on to accomplish that purpose, the set of assumptions

that underpin that capability, and the measurements that are both expected by that capability and actually produced by it. We adapt the socio-technical matrix to focus more generally on measurement systems and propose that contextual audits should use the qualitative methods described above to answer the following before conducting quantitative testing:

1. Identify the object(s) of capture and inference
2. Determine how objects are made legible to the system
3. Examine how ground truth is and has historically been established
4. Determine what assumptions “ground truths” encode
5. Consider how these assumptions may produce harms when operationalized in the relevant context of use

4 Case Study Audit: Skeleton Inference

Mocap systems operate in multiple contexts, ranging from entertainment to health and safety to biometric surveillance (Mündermann, Corazza, and Andriacchi 2006; Berner et al. 2020). Each context brings with it a unique and specific set of practices through which measurements are produced, and each practice is defined by its own set of materials, competences and meanings. Our case study audit focuses on the context of **mocap for creative design**, which entails capturing bodies in motion as inputs for creative projects (Dykes, Rodgers, and Smyth 2009). This can include faithfully recording body movement (e.g., for capturing an athlete’s signature style for a video game) as well as distorting bodies (e.g., for enabling CGI of nonhuman bodies performing natural-looking movements). When bodies are recorded, mocap systems perform skeleton inference and produce BSPs, which may then be used in downstream tasks.

Audit Stakes. Mocap for creative design is relatively low-stakes when compared to other mocap use cases, such as workplace safety monitoring, biometric surveillance, and medical diagnosis: a person’s physical health and safety are not at risk if they are mismeasured by a mocap system in a creative design task. Nevertheless, even a creative design mocap system has real stakes. In particular, mocap systems can contribute to representational harms if they misrepresent or erase a particular body type in entertainment settings (Blodgett et al. 2020). Further, as Harvey et al. (2024) show, materials, competences, or meanings initially developed for entertainment are often adapted for higher-stakes settings, meaning harms can travel with them. For example, the Microsoft Kinect mocap system was originally marketed as a video game controller, but quickly became the basis for home health and rehabilitation mocap applications due to its low price and small size (Blumrosen et al. 2016; Su, Chiang, and Huang 2014). Thus, while our audit focuses on mocap for creative design, its results point to areas that should be scrutinized in future audits of higher-stakes mocap systems.

Audit Setting. In our audit, we worked with mocap practitioners at New York University (NYU) whose work focused on researching augmented, virtual, and extended reality as well as working with external clients, including companies,

museums, and artists, to produce creative outputs.² Mocap at NYU is conducted on a *mocap stage*, a 40’x35’ platform that is surrounded by an OptiTrack Prime 13 mocap system, which includes 24 cameras arranged around the stage.³ The OptiTrack Prime 13 is a marker-based system: it uses infrared cameras to track markers placed on the surface of the human body. Those markers are then used to infer the position and motion of the underlying skeletal elements based on one of several *body models*, which are mathematical models built on historical data that come with OptiTrack’s Motive software. The body models perform skeleton inference and produce BSPs. Our study was approved by the NYU Institutional Review Board (IRB-FY2023-7677).

4.1 Incorporating Context

Qualitative Methods Reveal Context. To understand the context of mocap for creative design, we conducted field work directly with mocap practitioners at NYU in June and July 2023. We first spent a day in training. This included shadowing the practitioners as they conducted mocap and learning how to conduct mocap activities ourselves. We were trained on how to calibrate cameras,⁴ how to identify body landmarks and affix markers to them, how to capture motion using the OptiTrack system, and how to troubleshoot the system. Importantly, practitioners walked us through the actual practices that they followed as opposed to potential best practices that might exist in academic literature or system documentation. We also engaged practitioners in question-and-answer sessions to document their understandings of the mocap technology and the relevance of each of their techniques to the overall practice they were engaged in. Following the day of training, we conducted two (non-consecutive) weeks of field experiments using the mocap system, for which we recruited 24 participants to have their BSPs measured. During these experiments, we asked the mocap practitioners for guidance or help debugging the mocap system as needed. Finally, we conducted informal interviews with the mocap practitioners to better understand how they perceived the materials, competences, and meanings that were central to their practice of mocap. Throughout, we took extensive field notes to capture both the practices that practitioners described as well as how we applied and adapted those practices ourselves.

We supplemented our field work with a close examination of the OptiTrack documentation and a historical literature review (see Harvey et al. 2024). Together, the documentation and literature review allowed us to uncover materials, competences, and meanings that have been baked into the OptiTrack system over time, and thus help constitute the context in which the mocap practitioners we observed operate.

Context Reveals Assumptions. In our field work, we observed that practitioners considered valid representations of bodies to be those that were faithful to the proportions

²<https://engineering.nyu.edu/research/nyu-tandon-yard>

³<https://optitrack.com/>

⁴Calibration consists of waving a bar with markers attached at fixed points across the mocap stage to allow cameras to triangulate their positions relative to each other.

	Marker-Based Motion Capture	Tape Measure-Based Anthropometry
<i>Identify the object(s) of capture and inference</i>	<i>Capture:</i> Markers on bodies <i>Inference:</i> Position and motion of skeleton	<i>Capture:</i> Distance between landmarks on bodies <i>Inference:</i> Body segments
<i>Determine how these objects are made legible to the system</i>	System inputs are markers that are placed on body landmarks as specified by a body model. Multiple calibrated cameras sense each marker, which allows the XYZ position of markers to be inferred. Then, the body model is then used to infer the position and motion of the underlying skeleton.	System inputs are body landmarks identifiable from the surface of the body, which a tape measure is used to measure the distance between. Both the identification of landmarks and the application of tape measures are informed by public health guides to anthropometric measurement.
<i>Examine how ground truth is and has historically been established</i>	<i>Historically:</i> The earliest ground truth data on BSPs used in motion capture relied on dismembering dead bodies or implanting pins into the bones of living bodies in an attempt to measure BSPs directly. Since then, researchers have relied primarily on concurrent validity assessments which show that new mocap systems produce measurements similar to existing systems (Harvey et al. 2024). <i>Current Context:</i> Our audit followed the ground truthing practices of the mocap practitioners at our field site: ball-and-stick camera calibration, review of marker placement by multiple researchers, and concurrent recording of motion with markers, video cameras, and a depth camera.	<i>Historically:</i> Units of measurement were socially constructed (e.g., based on easy-to-measure and relevant lengths, like armspan) and then standardized (e.g., by measuring how far light travels in a specific fraction of a second). Anthropometric measurement landmarks were originally chosen to study variation in physical anthropology, and later, to aid biometric identification in criminal justice and for eugenic purposes (Maguire 2009). <i>Current Context:</i> Our audit followed anthropometric guides produced by the U.S. government (DHHS 1994) as well as analogous ground-truthing practices to those used for mocap, including review of body landmark identification by multiple researchers and concurrent validity assessments of measurement instruments (i.e., comparing tape measures to one another, measuring a known weight on the scale).
<i>Determine what assumptions “ground truths” encode</i>	People have standard body landmarks (corresponding to those present on the bodies of primarily male research subjects) and have a standard amount of soft tissue between markers and the surface of bodies (Keller et al. 2023).	People have standard body landmarks and body landmarks are equally identifiable on people regardless of their body shape or size.
<i>Consider how these assumptions may produce harms when operationalized in the relevant context of use</i>	We expect to see less fidelity of representation for people who are not male or not thin. This may cause a person to see their body (or a body like theirs) misrepresented or not represented in entertainment. Measurement processes initially used for entertainment may be adapted for higher-stakes settings, and harms can travel with them.	We expect to see less fidelity of representation for people who are not close to the average used to construct anthropometric measurement procedures and norms. Measurements indicating that individuals are further from the average may be used to reify or reinforce eugenicist harms.

Table 1: The **Contextual Auditing Matrix**, completed for the context of mocap for creative design and considering both a marker-based motion capture system and a tape measure.

and movements of subjects on the mocap stage. To produce measurements seen as valid, practitioners attended to the calibration of cameras, the proper placement of markers on subjects’ bodies, the management of reflective objects on the stage that might interfere with mocap recordings, and areas of the mocap stage with greater and lesser accuracy and reliability (due to camera placement). We observed that, while practitioners undertook practices to produce what they would see as valid measurements, they ultimately took validity as an article of faith. They judged when a set of measurements “look[ed] right” for assurances that their systems were working properly. They looked for areas of the mocap stage where measurements “glo[t] wonky” to decide when to recalibrate. By “wonky,” practitioners meant that the dimensions and positions of skeletal elements fell out of proportion

in the mocap software—elongated necks, twisted limbs, and foreshortened arms were clues practitioners used to identify a need to recalibrate the system. However, in the absence of egregious distortions, they trusted that the calibrated mocap system produced accurate measurements rather than verifying those measurements themselves (by, for example, measuring subjects with an alternate apparatus).

Our field work thus directed us towards key assumptions encoded in the practice of mocap for creative design and disclosed how ground truth was established; namely that placing markers consistently and calibrating the mocap stage correctly produces valid measurements for any human subject, and that, when they occur, invalid measurements can be identified heuristically. To interrogate these assumptions, we chose to directly audit a basic set of measurements produced

by the mocap system: BSPs (specifically, segment length).

Assumptions Reveal Facets. BSPs are an appropriate basis for a contextual audit of a mocap system because it is easy to reason about what they should look like: the BSPs of an individual should be stable over a short period of time, and estimates of BSPs produced by a mocap system should be similar to estimates of BSPs produced by alternative measurement modalities. Based on the OptiTrack documentation, analyzed through the lens of our previously conducted literature review (Harvey et al. 2024), we identified several facets across which to audit the mocap system. The system should perform equally well (e.g., produce equally stable measurements) across the facets of *body size*, *sex*, and *time*.

We chose these facets based on the assumptions encoded in marker-based mocap systems. As Harvey et al. (2024) outline, and as our review of the OptiTrack documentation confirms, mocap systems were built using data from a small number of bodies that were overwhelmingly male and thin. This means that different types of bodies may be more likely to be misrepresented by mocap systems. We chose to focus on the facets of body size and sex in particular because the mocap system we worked with was a marker-based system, and body size and sex are both likely to affect the location of body landmarks (and, correspondingly, marker placement). We also chose to include time as a facet in order to capture variation related to the practice of body measurement using mocap (e.g., room calibration or marker placement on the human subject by experimenters) as well as potential variation in the human subject (e.g., weight gain, change in posture due to injury).

4.2 Incorporating Symmetry

There is no single agreed-upon system of “ground truthing” BSPs. For example, consider the distance between a person’s shoulders. One could measure the length of the collarbone, or the straight-line distance between shoulder blades. One could determine the length based only on bones (which are not directly observable while the bones are contained within the body), or one could take soft tissue into account. Universal ground truth values for BSPs are unknown—and potentially unknowable. Thus, we did not have access to a definitive ground truth against which to compare measurements produced by the mocap system. Instead, following practices common in concurrent validity assessments, we compared BSPs produced by the mocap system to anthropometric measurements produced using a tape measure. We chose this alternative modality because it is free of many of the starting assumptions of mocap technology and was easily accessible to us (unlike, e.g., an X-ray machine). Our approach differed from a typical concurrent validity assessment because we incorporated the concept of symmetry. That is, we built upon the understanding that neither the tape measure nor the mocap system could or should be taken as a ground truth. Rather than interrogate the OptiTrack mocap system by taking tape measured BSPs as a ground truth, we used the framework of contextual auditing to examine the assumptions underlying *both* systems. This analysis is summarized in Table 1. Then,

in our empirical analysis, we built on the notion of symmetry to compare how the different measurements varied with respect to one another, but did not characterize this as systematic bias from a ground truth. In this way, incorporating the concept of symmetry allowed us to conduct a contextual AI audit even in the absence of an observable ground truth.

4.3 Audit Protocol

After developing an understanding of the context and assumptions surrounding mocap for creative design, we constructed an audit protocol to test the stability of BSPs across our identified facets. The experimental protocol consisted of two measurement phases. In Phase 1, participants’ measurements were recorded using traditional anthropometric techniques and equipment (DHHS 1994). In Phase 2, participants’ measurements were recorded using the mocap system. To conduct a contextually grounded audit, we embedded ourselves within the mocap environment, strictly adhering to the laboratory’s local practices. While the fundamental procedures align with the official OptiTrack Motive documentation,⁵ the specific workflows were executed according to internal facility protocols to ensure consistency with the site’s standard operating procedures.

The protocol begins with participants’ arrival at the field site, at which time they were asked to provide informed consent to participate, given the opportunity to ask questions about the study, and assigned a participant ID to separate the collected data from the personal information used to manage their enrollment. Participants were then issued a mocap suit—a form-fitting spandex bodysuit with a Velcro exterior appropriate for affixing the markers used in mocap. The mocap suit also facilitated the collection of body measurements in Phase 1, as it removed any obstructions that might have been introduced by structured or bulky clothing.

Tape Measure. For Phase 1, each set of body measurements was taken by two members of the research team. In each pair, one researcher took the role of primary measurer, while the other recorded measurements and assisted in measuring wingspan (a two-person job). All measurements except for weight were read aloud by the primary measurer and then verbally confirmed and recorded by the assistant measurer. In most, but not all, cases, we matched the sex of the primary measurer to the sex of the participant that they were measuring. Body measurement protocols were drawn from the National Health and Nutrition Examination Survey (DHHS 1994). All measurements were taken using a soft tape measure, except for weight, which was measured with a digital scale. The specific measurements recorded, listed in the order in which they were taken, were: standing height, wingspan, sitting height, upper leg length, biacromial breadth, upper arm length, upper arm circumference, abdominal circumference, hip circumference, thigh circumference, head circumference, and weight. The full measurement protocol is available in Appendix A.

Mocap System. One researcher (the technical lead) performed daily room calibration in accordance with standard

⁵<https://docs.optitrack.com/motive/skeleton-tracking>

procedures. This involved clearing the mocap stage of reflective noise and performing a wand wave calibration using the OptiTrack large space calibration wand (a rigid pole with markers affixed to known locations). The wand was moved in a figure-8 pattern until a minimum of 10,000 samples were registered per camera. The team adhered to a strict error threshold, proceeding only when the system reported an “Exceptional” rating (approximate mean error ≤ 0.6 mm). The ground plane was defined using an L-shaped calibration square aligned with the participants’ walking direction.

Participants were fitted with a standard mocap suit (cap, gloves, and shoes) secured with Velcro. The research team attached 41 retroreflective markers to the suit following the standard OptiTrack “Baseline 41” body model. One of the mocap practitioners at the facility, a specialist with extensive experience in the entertainment industry, performed periodic quality control checks to ensure marker placement aligned with industry standards. Following marker placement, the technical lead conducted subject calibration. Participants assumed a static T-pose (standing with feet shoulder-width apart, arms raised to 90 degrees) to allow the Motive software to infer the body model automatically. Neither manual alignment of the skeleton nor dynamic calibration was performed. The technical lead then assigned a generic avatar for real-time visualization purposes.

Once calibrated, participants followed a scripted protocol of distinct everyday movements: performing a T-pose, walking, sitting, rotating, squatting, swinging arms, making circles with arms, jumping, kicking, carrying a 10-lb kettlebell (in front and to the side of the body), carrying a large exercise ball, stepping onto a platform, bouncing a kickball, cross toe touching, running in place, using a cell phone, pointing, checking the time on a watch, rubbing hands together, crossing arms, and planking. These movements were chosen because they represent everyday body movements that may feature in entertainment products using mocap, as well as actions that might induce errors in mocap systems (as they may cause markers to be occluded or change a participant’s center of gravity). While the technical lead operated the Motive software and recording from the control desk, two additional researchers read out the movement protocol and demonstrated movements where appropriate. We recorded the mocap data produced via the OptiTrack system as well as auxiliary data; namely, a depth camera recording of the mocap stage (OAK-D) and a screen capture of the live Motive interface. To export BSPs from the system, we exported the Motive recording as a CSV containing, for every frame, the 3D position and orientation quaternion of each skeletal element produced by the Baseline 41 body model. From these files we read the endpoints of each BSP as the Euclidean distance between two endpoints: shoulder width as left to right upper arm; upper arm length as upper arm to forearm; lower arm length as forearm to hand; upper leg length as thigh to shin; lower leg length as shin to foot; and standing height as the mean of the head-to-foot distances on the left and right sides. Distances were computed on the first 10 frames of each take, captured while the participant held the static T-pose used for subject calibration, and averaged to yield a single BSP value per measurement per session.

The extraction script is available in our public repository⁶ and the full mocap protocol is available in Appendix B.

4.4 Data Analysis

Our analysis focuses on a set of BSPs: the estimates of standing and sitting height, wingspan, biacromial breadth (distance between shoulders) and upper arm and leg length produced by OptiTrack and by tape measurements. One intuitively desirable characteristic of BSP measurement is that it should be reliable. In other words, the BSPs of an individual should be stable over a short period of time or when produced by different measurement instruments. Therefore, we focus on the difference between measurements—e.g., across measurement techniques or over time—as the dependent variable in our analysis. The independent variables are the facets of body size (operationalized as BMI, weight and body segment size),⁷ sex (self-reported), and time (first or second session), as well as measurement modality (mocap vs. tape measure). Within each session, we compare measurements of the same body part across the different modalities. We also compare measurements of the same body part across different sessions. We systematically compare how reliable our measurements are with respect to facets of interest. Noting that our results are fundamentally exploratory, not confirmatory, we test the following hypotheses:

- H0** Mocap measurements for the same participant are reliable across time points.
- H1** The relationship between tape measured and mocap measurements is reliable across participant body size.
- H2** The relationship between tape measured and mocap measurements is reliable across participant sex.
- H3** The relationship between tape measured and mocap measurements is reliable across time points.

Statistical Models. We build upon a statistical model commonly used in mocap studies and reliability engineering more broadly: the Bland-Altman limits of agreement (LOA) (Altman and Bland 1983). We extend the Bland-Altman LOA with nested regression models to explore the relationship between reliability and facets of interest. Calculating the Bland-Altman LOA involves regressing the difference between paired measurements on the average of the paired measurement. We focus specifically on comparisons of measurements taken of the same body part of the same subject. To determine whether a measurement should be considered reliable across a modality or facet of

⁶https://github.com/HaukeCornell/Audit_Skeletons

⁷Body Mass Index, or BMI, is a ratio of weight to height ($\frac{\text{weight (kg)}}{\text{height (m)}^2}$). BMI was originally developed for white male bodies and intended as a population-level statistic (Pray and Riskin 2023). Today, BMI is applied to bodies of all types as a measure of individual health. Although BMI is problematic, it is correlated with the presence of soft tissue (i.e., skin, fat, and muscle), and soft tissue is associated with mocap error (Jeong et al. 2023). BMI is less invasive than other approaches for measuring soft tissue, such as skinfold calipers. We supplement BMI with weight and body segment size in order to avoid relying on BMI alone.

interest, we define reliability as having regression coefficients that are not statistically significantly different from zero ($p \geq 0.05$, corrected for multiple comparisons). If a measurement is reliable, measurement differences do not vary with respect to the included features.

Across-session, we compare across measurements produced by modality x for individual i and body part j in Sessions 1 and 2. We apply the following model, iteratively adding explanatory variables:

$$x_{ij1} - x_{ij2} \sim \frac{x_{ij1} + x_{ij2}}{2} + \text{BMI}_{it} + \text{weight}_{it} + \text{sex}_{it}$$

Within-session, we compare across measurement modality x and y , for individual i and body part j in session t . We apply the following model, again iteratively adding explanatory variables:

$$x_{ijt} - y_{ijt} \sim \frac{x_{ijt} + y_{ijt}}{2} + \text{session}_t + \text{BMI}_{it} + \text{weight}_{it} + \text{sex}_{it}$$

Following typical practice for Bland-Altman LOA, a nonzero coefficient for the first term would indicate systematic variation across modalities with respect to the target of measurement. The other variables (session, weight, sex, etc.) would capture systematic variation in measurement reliability across methods with respect to those facets, all else constant. We pre-registered our plan prior to conducting data analysis.⁸ Our analysis code is described and linked in Appendix C.

4.5 Audit Results

We applied the protocol to 24 unique participants, 12 of whom were measured during two sessions approximately one month apart. The participants ranged in age (21–63, median: 32), height (153–193 cm, median: 171 cm), and weight (45.5–115.5 kg, median: 71 kg), and sex (male: 11, female: 13). Because of the large number of variables and consistent findings, we briefly summarize our key results here. Our full results are available in Appendix D.

H0: *Mocap measurements for the same participant are reliable across time points.* This mostly holds, except for wingspan and biacromial breadth, where the difference increases as the average measurement increases.

H1: *The relationship between tape measured and mocap measurements is reliable across participant body size.* This mostly holds, with two exceptions. First, the difference between biacromial breadth as measured by a tape measure vs. OptiTrack increases as average biacromial breadth decreases. This holds whether or not time, BMI, weight, or sex are accounted for. Second, for standing height, the difference increases as the average measurement increases. Conditional on average standing height, this deviation increases as BMI increases. Conditional on those, it decreases as weight increases. This holds whether or not sex is accounted for.

H2: *The relationship between tape measured and mocap measurements is reliable across participant sex.* This holds (i.e., no statistically significant differences).

H3: *The relationship between tape measured and mocap measurements is reliable across time points.* This holds.

⁸<https://bit.ly/osfxy4jg>

We emphasize that our results are fundamentally *exploratory* and should only be interpreted as such. Any detected systematic differences across facets should be seen as hypothesis-generating for future analysis, not evidence of explicit bias of either the OptiTrack mocap system or of a tape measure.

5 Discussion

The contextual audit framework allows auditors to explore measurements in the context of the practices that produced them. In this way, contextual audits provide a way of answering the two fundamental questions raised by AI auditing: “what inputs are relevant for identifying patterns in a system’s actual outputs?” and “what should a system’s nominal outputs look like?” In our case study audit (§4), the contextual framework enabled us to identify assumptions underlying BSP measurement as conducted by mocap systems as well as facets across which to audit. By incorporating the concept of symmetry from STS, we were able to conduct a contextual audit of mocap’s ability to measure BSPs, despite the fact that there is no single agreed-upon approach for “ground-truthing” BSPs—that is, in the context of BSPs, ground truth can be constructed or approximated in multiple different ways. We did this by using an alternate measurement modality to serve as a provisional ground truth and treating that provisional ground truth as the nominal outputs according to which we evaluated actual system outputs. In turn, that modality could also be evaluated using mocap: under the principle of symmetry, each measurement modality can serve as ground truth for the other.

The value of symmetry in auditing skeleton inference is illustrated by **H1** above, where the actual and nominal outcomes vary in two cases. First, the variation between mocap and tape measured estimates of biacromial breadth increases as segment length decreases. Applying the principle of symmetry, we can explain this variation from the position that either modality is potentially true while the other is not. The tape measurements may be particularly sensitive to subjects’ posture when measuring across their shoulders, *or* there may be systematic irregularities having to do with shoulder width in the mocap system.⁹ Second, the variation between mocap and tape measured estimates of height increases as height increases. There could be systematic error in the anthropometric protocol used to measure standing height—perhaps the protocols are poorly calibrated in ways that are only apparent when measuring larger values (standing height being the largest set of values measured with the tape measure). Symmetrically, there could be systematic error in the mocap system, as suggested by Harvey et al. (2024), for taller individuals with higher BMIs, who were under-represented in the original BSP calculations. The audit does not resolve the question of what causes these unreliable measurements. Rather, applying symmetry demonstrates how audit results can still be used to surface practical next steps for the audit process when ground truth is unknown, unknowable, or contested.

⁹Or tape measurements may be applied inconsistently across researchers, or some combination of these and other processes.

Audit Implications. As discussed in §2.1, a primary goal of AI auditing is accountability (Birhane et al. 2024; Bovens 2007). Thus, we make our expectations of accountability explicit here. As an exploratory audit, we do not directly identify disparities that mocap providers must “fix” (although future confirmatory audits may do so). Instead, we call on both mocap providers and researchers to consider the assumptions underlying mocap systems in their work; to test system performance across the relevant facets of sex and body size; and to go beyond straightforward concurrent validity assessments in their analyses. More broadly, we call on researchers and practitioners to consider ecological validity by using qualitative methods to understand the context in which systems are actually designed and used before conducting AI audits.

Future Work. We identify several opportunities for future work. First, we recognize the importance of building on our case study audit with additional studies designed to resolve the questions raised by our symmetrical approach. Second, we call for contextual audits to be adopted as an approach for auditing other AI systems. We argue that it is particularly urgent to adopt a contextual audit approach for settings where measurement practices carry heavy traditions of dehumanizing, pseudoscientific, and otherwise harmful assumptions (e.g., in hiring, where AI/ML-based tools have been shown to have eugenicist and pseudoscientific roots (Sloane, Moss, and Chowdhury 2022)). AI audits that fail to adopt a contextual approach (e.g., hiring audits that do not assess tools in their actual contexts of use) may instead perpetuate the harms they intend to mitigate.

While we emphasize that our case study audit is exploratory, we can still identify opportunities for future audits into mocap systems, especially in higher-stakes settings. We suggest that AI audits should focus on facets that displayed lower stability in our case study audit by researching how well mocap systems perform for bodies with narrower shoulders as well as for larger bodies. This is particularly important in the context of prior work by Harvey et al. (2024), who found that early research into mocap systems, which has stabilized into current practices, overwhelmingly included thin bodies—for example, just 5% of subjects in early research on soft tissue artifacts, which occur when skin, muscle, and fat cause markers on the surface of subjects’ bodies to move independently of subjects’ bones, were noted as “overweight” (according to BMI). Modern mocap research acknowledges, but still does not address, that mocap methods are often invalid for subjects with more soft tissue (Chu et al. 2012); it may attribute errors in mocap systems “to the anatomy of the [subject]” and not necessarily to the mocap system itself (Paloschi et al. 2021). Thus, future mocap audits with a continued focus on body size as a facet are critical.

Finally, we call for additional research into how to conduct audits when ground truth is unknown, unknowable, or contested. While the concept of symmetry can help identify areas of instability for future audits, as we show in our study, confirmatory audits may still require a more explicit ground truth.

Limitations. The case study audit we describe in §4 is intended to illustrate our conceptual contribution: a rigorous methodology to incorporate the social practice focus of contextual audits into the statistical framework of AI auditing. The results we share are fundamentally exploratory and should not be taken as a definitive endorsement or critique of either OptiTrack or tape measures.

Our experimental setup highlights many of the idiosyncratic choices that become encoded in typical AI audits. The mocap setting specifically requires choices about types of populations (demographics, ranges of body types, ranges of mobility and mobility aids, etc.) that are already poorly cataloged in existing mocap and computer vision development and validation practices (Sloane, Jacobs, and Moss 2025). Our study, for example, did not include participants who used mobility aids and primarily (although not deliberately) included participants who fell within the so-called ‘normal’ BMI category. In addition, our small sample size, non-expert measurers, and multiple regressions limit the empirical claims we can make with our data.

Perhaps most importantly, we note that we are researchers, and not truly real-world system users, i.e., expert practitioners with years of experience in mocap for creative design. As Harvey et al. (2024) describe, errors in measurements produced by mocap systems have several sources: while some may be related to the mocap system itself and the assumptions it makes, others are attributable to the operators of a mocap system. In particular, operators produce *marker placement errors*, which occur when the markers used by a mocap system are not placed correctly on the body landmark specified by a body model (Harvey et al. 2024). We anticipated that the magnitude of marker placement error produced by us, as non-expert mocap users, would be larger than the magnitude of marker placement error produced by the expert practitioners whose practices we sought to replicate. This complicates how audit results can be interpreted: rather than an interrogation of the system alone, our audit produces an interrogation of the system and its operators (us). We believe that this is likely to be broadly true for audit studies in which researchers assess a system typically used by experts, and call for contextual audits to explicitly include a *reflexive* step in which auditors consider how to disentangle what audit results say about AI systems from what audit results say about auditors.

6 Conclusion

We introduce **contextual audits** as a method for auditing measurements within the context of the practices that produce them. We show how incorporating the concept of **symmetry** from STS can enable auditors to be explicit about what serves as a ground truth and, if needed, conduct audits when ground truth is unknown, unknowable, or contested. Finally, to demonstrate how contextual and symmetric audits can be conducted in practice, we present a case study of skeleton inference in motion capture.

Ethical Statement

Researcher Positionality. We are a multidisciplinary team of researchers, with academic backgrounds in computer science, data science, information science, anthropology, and sociology. Our backgrounds have shaped the disciplinary lenses that we applied to our audit, leading us to draw on the concepts of social practice theory from sociology and reflexivity and symmetry from STS. Our backgrounds have also shaped the questions we asked in our research. While we study mocap systems, no member of the research team has professional experience as a mocap practitioner. Several members of the team, however, have experience in the arts or in creative design. Our bodies, in some ways, look like those that mocap systems were originally designed for (we are all white and cisgender); in other ways, they do not (more than half of us are women). We thus approached our audit as outsiders to mocap for creative design, but with the lived experience of seeing bodies that looked like ours represented (and sometimes misrepresented) in mocap-enabled visual arts.

Ethical Considerations. The primary ethical considerations we made for this study were related to the safety and privacy of audit participants. Our study was approved by a university Institutional Review Board. All participants provided written informed consent prior to engaging in the study. Participants were compensated for their time with a \$15 gift card per session attended, an amount set to ensure participants would not feel compelled to participate due to financial need. At the conclusion of the session, all participants were given the opportunity to participate in a semi-structured ethnographic interview regarding their experience of being measured and digitized. Finally, in order to protect participant privacy, all study data was stored separately from participant names using anonymized ID variables and all video data was de-identified using facial blurring. An additional ethical concern relates to the historical uses of anthropometry for race science and eugenics (Clever 2023; Turda 2010). Our implementation of anthropometric techniques was strictly limited to vetted, contemporary methods widely deployed in healthcare (DHHS 1994).

Adverse Impact. We believe that our study poses only minimal risks to participants. Participants were asked only to conduct everyday movements, such as stepping and bouncing a ball, and were explicitly instructed not to complete movements that felt uncomfortable to them. To mitigate potential emotional harms associated with being weighed, we gave all participants the option of being weighed while facing away from the scale display, and did not read participant weights aloud. Participants also faced the risk of emotional harm from seeing their bodies misrepresented in the mocap software; however, based on exit interviews, we believe that participants did not experience this as a harm.

Like any audit of an AI system, our study runs the broader risk of inadvertently contributing to improving or legitimizing technology that enables large-scale surveillance or other negative societal effects. We believe that in our case, this risk is minimal, as the audit we present is intended to be an illustrative case study of contextual auditing as opposed to a rigorous audit of the OptiTrack mocap system.

Acknowledgments

This work was supported by funding from the Notre Dame-IBM Technology Ethics Lab. We thank Todd Bryant, Luke DuBois, Kaustav Sarkar, and Harsh Palan for their valuable support.

References

- Altman, D. G.; and Bland, J. M. 1983. Measurement in Medicine: The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(3): 307–317.
- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. ProPublica.
- Bandalos, D. L. 2018. *Measurement theory and applications for the social sciences*. Guilford Publications.
- Bandy, J. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 74:1–74:34.
- Becerra Sandoval, J. C.; and Jing, F. S. 2025. Historical Methods for AI Evaluations, Assessments, and Audits. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1371–1386. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Berner, K.; Cockcroft, J.; Morris, L. D.; and Louw, Q. 2020. Concurrent validity and within-session reliability of gait kinematics measured using an inertial motion capture system with repeated calibration. *Journal of Bodywork and Movement Therapies*, 24(4): 251–260.
- Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4): 991–1013.
- Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 612–643. Los Alamitos, CA, USA: IEEE Computer Society.
- Blodgett, S. L.; Barocas, S.; Daumé Iii, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, 5454–5476. Online: Association for Computational Linguistics.
- Bloor, D. 1991. *Knowledge and Social Imagery*. Chicago, IL: University of Chicago Press, 2 edition.
- Blumrosen, G.; Miron, Y.; Intrator, N.; and Plotnik, M. 2016. A real-time kinect signature-based patient home monitoring system. *Sensors*, 16(11): 1965.
- Bodenheimer, B.; Rose, C.; Rosenthal, S.; and Pella, J. 1997. The process of motion capture: Dealing with the data. In *Computer Animation and Simulation'97: Proceedings of the Eurographics Workshop in Budapest, Hungary, September 2–3, 1997*, 3–18. Springer.

- Bovens, M. 2007. Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, 13(4): 447–468.
- Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, FAT*, 77–91. PMLR.
- Cheng, H.-F.; Stapleton, L.; Kawakami, A.; Sivaraman, V.; Cheng, Y.; Qing, D.; Perer, A.; Holstein, K.; Wu, Z. S.; and Zhu, H. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*, CHI, 1–22. New Orleans LA USA: ACM. ISBN 978-1-4503-9157-3.
- Chu, Y.; Akins, J.; Lovalekar, M.; Tashman, S.; Lephart, S.; and Sell, T. 2012. Validation of a video-based motion analysis technique in 3-D dynamic scapular kinematic measurements. *Journal of Biomechanics*, 45(14): 2462–2466.
- Clever, I. 2023. Biometry against fascism: Geoffrey Morant, race, and anti-racism in twentieth-century physical anthropology. *Isis*, 114(1): 25–49.
- Durkin, J. L.; and Dowling, J. J. 2003. Analysis of Body Segment Parameter Differences Between Four Human Populations and the Estimation Errors of Four Popular Mathematical Models. *Journal of Biomechanical Engineering*, 125(4): 515–522.
- Dykes, T. H.; Rodgers, P. A.; and Smyth, M. 2009. Towards a new disciplinary framework for contemporary creative design practice. *CoDesign*, 5(2): 99–116.
- Feng, Y.; and Shah, C. 2022. Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 11882–11890.
- Fleisig, E.; Abebe, R.; and Klein, D. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 6715–6726. Singapore: Association for Computational Linguistics.
- Fleisig, E.; Smith, G.; Bossi, M.; Rustagi, I.; Yin, X.; and Klein, D. 2024. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 13541–13564. Miami, Florida, USA: Association for Computational Linguistics.
- Fuller, S. 2016. Embrace the inner fox: Post-truth as the STS symmetry principle universalized. *Social Epistemology Review and Reply Collective*, 25.
- Geiger, R. S.; and Ribes, D. 2011. Trace Ethnography: Following Coordination through Documentary Practices. In *2011 44th Hawaii International Conference on System Sciences*, 1–10.
- Gosciak, J.; Giannella, E.; Guo, Z.; Chen, M.; and Koenecke, A. 2026. LLMs in social services: How does chatbot accuracy affect human accuracy? arXiv:2603.11213.
- Han, S.; and Lee, S. 2013. A vision-based motion capture and recognition framework for behavior-based safety management. *Automation in Construction*, 35: 131–141.
- Hand, D. 2004. *Measurement Theory and Practice: The World Through Quantification*. Wiley. ISBN 9780340677834.
- Harvey, E.; Kizilcec, R. F.; and Koenecke, A. 2025. A Framework for Auditing Chatbots for Dialect-Based Quality-of-Service Harms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, 2025–2039. Athens Greece: ACM. ISBN 979-8-4007-1482-5.
- Harvey, E.; Sandhaus, H.; Jacobs, A. Z.; Moss, E.; and Sloane, M. 2024. *The Cadaver in the Machine: The Social Practices of Measurement and Validation in Motion Capture Technology*. *Proc. ACM CHI*.
- Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028): 147–154.
- Imana, B.; Shen, Z.; Heidemann, J.; and Korolova, A. 2025. External Evaluation of Discrimination Mitigation Efforts in Meta’s Ad Delivery. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, 2616–2629. Athens Greece: ACM. ISBN 979-8-4007-1482-5.
- Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, 375–385. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8309-7.
- Jasanoff, S. 2019. Controversy studies. *The Blackwell encyclopedia of sociology*, 1–5.
- Jeong, S.-M.; Lee, D. H.; Rezende, L. F. M.; and Giovannucci, E. L. 2023. Different correlation of body mass index with body fatness and obesity-related biomarker according to age, sex and race-ethnicity. *Scientific Reports*, 13(1): 3472.
- Keller, M.; Werling, K.; Shin, S.; Delp, S.; Pujades, S.; Liu, C. K.; and Black, M. J. 2023. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *ACM Transactions on Graphics (TOG)*, 42(6): 1–12.
- Kihlstrom, J. F. 2021. Ecological Validity and “Ecological Validity”. *Perspectives on Psychological Science*, 16(2): 466–471. PMID: 33593121.
- Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A. M.; Margatina, K.; Mosquera, R.; Ciro, J. M.; Bartolo, M.; Williams, A.; He, H.; Vidgen, B.; and Hale, S. A. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Klein, H. 1988. *The Science of Measurement: A Historical Survey*. Dover Science Books. Dover Publications. ISBN 9780486258393.
- Landers, R. N.; and Sanchez, D. R. 2022. Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. *International Journal of Selection and Assessment*, 30(1): 1–13.

- Maguire, M. 2009. The birth of biometric security. *Anthropology today*, 25(2): 9–14.
- Mei, K. X.; Choi, A. S. G.; Schellmann, H.; Sloane, M.; and Koenecke, A. 2026. Addressing Auditing Pitfalls in Automatic Speech Recognition Technologies: A Case Study of People with Aphasia. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, 3422–3465. New York, NY, USA: Association for Computing Machinery. ISBN 9798400725968.
- Mökander, J. 2023. Auditing of AI: Legal, ethical and technical approaches. *Digital Society*, 2(3): 49.
- Mündermann, L.; Corazza, S.; and Andriacchi, T. P. 2006. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of neuroengineering and rehabilitation*, 3(1): 6.
- Mussnug, A. M. 2022. The predictive reframing of machine learning applications: good predictions and bad measurements. *European Journal for Philosophy of Science*, 12(3): 55.
- Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.
- Paloschi, D.; Bravi, M.; Schena, E.; Miccinilli, S.; Morrone, M.; Sterzi, S.; Saccomandi, P.; and Massaroni, C. 2021. Validation and Assessment of a Posture Measurement System with Magneto-Inertial Measurement Units. *Sensors*, 21(19).
- Pray, R.; and Riskin, S. 2023. The history and faults of the body mass index and where to look next: A literature review. *Cureus*, 15(11): e48230.
- Pruss, D. 2023. Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, 312–323. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0192-4.
- Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 959–972. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. 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. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.
- Rhea, A.; Markey, K.; D'Arinzo, L.; Schellmann, H.; Sloane, M.; Squires, P.; and Stoyanovich, J. 2022a. Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring: Results of an Audit. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES, 572–587. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9247-1.
- Rhea, A. K.; Markey, K.; D'Arinzo, L.; Schellmann, H.; Sloane, M.; Squires, P.; Arif Khan, F.; and Stoyanovich, J. 2022b. An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Mining and Knowledge Discovery*, 36(6): 2153–2193.
- Rismani, S.; Shelby, R.; Smart, A.; Jatho, E.; Kroll, J.; Moon, A.; and Rostamzadeh, N. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.
- Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *"Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association*. Seattle, WA.
- Shove, E., ed. 2007. *The design of everyday life*. Cultures of consumption series. New York, NY: Berg. ISBN 9781845206826 9781845206833.
- Shove, E.; Pantzar, M.; and Watson, M. 2012. *The dynamics of social practice: everyday life and how it changes*. Los Angeles: SAGE. ISBN 9780857020420 9780857020437. OCLC: ocn783129226.
- Sloane, M. 2021. The Algorithmic Auditing Trap.
- Sloane, M.; Jacobs, A. Z.; and Moss, E. 2025. Making Bodies: Assumptions in the Design and Validation of Motion Capture Technology. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 2410–2418.
- Sloane, M.; and Moss, E. 2022. Introducing a Practice-based Compliance Framework for Addressing New Regulatory Challenges in the AI Field. *TechReg Chronicle*.
- Sloane, M.; and Moss, E. 2023. Assessing the Assessment: Comparing Algorithmic Impact Assessments and AI Audits.
- Sloane, M.; Moss, E.; and Chowdhury, R. 2022. A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, 3(2): 100425.
- Sloane, M.; Solano-Kamaiko, I. R.; Yuan, J.; Dasgupta, A.; and Stoyanovich, J. 2023. Introducing contextual transparency for automated decision systems. *Nature Machine Intelligence*, 5(3): 187–195.
- Su, C.-J.; Chiang, C.-Y.; and Huang, J.-Y. 2014. Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic. *Applied Soft Computing*, 22: 652–666.
- Sweeney, L. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5): 44–54.
- Topham, L. K.; Khan, W.; Al-Jumeily, D.; and Hussain, A. 2022. Human Body Pose Estimation for Gait Identification: A Comprehensive Survey of Datasets and Models. *ACM Comput. Surv.*, 55(6).
- Turda, M. 2010. Race, science, and eugenics in the twentieth century. *The Oxford handbook of the history of eugenics*, 62–79.

U.S. Department of Health and Human Services Public Health Service Centers for Disease Control and Prevention National Center for Health Statistics. 1994. Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. Series 1: programs and collection procedures.

Vecchione, B.; Levy, K.; and Barocas, S. 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO, 1–9. – NY USA: ACM. ISBN 978-1-4503-8553-4.

Wang, A.; Ho, D. E.; and Koyejo, S. 2025. The inadequacy of offline large language model evaluations: A need to account for personalization in model behavior. *Patterns*, 6(12).

Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; and Polli, F. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 666–677. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Wilson, K.; and Caliskan, A. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES2024)*, AIES.

Xsens. 2024. Motion Capture and the Future of Occupational Health and Safety.