# Teacher, Trainer, Counsel, Spy: How Generative AI can Bridge or Widen the Gaps in Worker-Centric Digital Phenotyping of Wellbeing

Vedant Das Swain
Northeastern University
Boston, MA, USA
v.dasswain@northeastern.edu

Koustuv Saha
University of Illinois Urbana-Champaign
Urbana, IL, USA
ksaha2@illinois.edu

## ABSTRACT

The increasing integration of computing technologies in the workplace has also seen the conceptualization and development of data-driven and algorithmic tools that aim to improve workers' wellbeing and performance. However, both research and practice have revealed several gaps in the effectiveness and deployment of these tools. Meanwhile, the recent advances in generative AI have highlighted the tremendous capabilities of large language models (LLMs) in processing large volumes of data in producing human-interactive natural language content. This paper explores the opportunities for LLMs in facilitating worker-centered design for Wellbeing Assessment Tools (WATs). In particular, we map features of LLMs against known challenges of WAT. We highlight how the LLMs can bridge or even widen the gaps in worker-centeric WAT. This paper aims to inspire new research directions focused on empowering workers and anticipating harms in integrating LLMs with workplace technologies.

## CCS CONCEPTS

• **Applied computing** → *Law, social and behavioral sciences*; *Psychology*; • **Human-centered computing** → *Empirical studies in collaborative and social computing*; *Social media.*

## KEYWORDS

workplace, generative AI, large language models, LLMs, worker wellbeing, worker performance

**ACM Reference Format:**
Vedant Das Swain and Koustuv Saha. 2024. Teacher, Trainer, Counsel, Spy: How Generative AI can Bridge or Widen the Gaps in Worker-Centric Digital Phenotyping of Wellbeing. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '24), June 25–27, 2024, Newcastle upon Tyne, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3663384.3663401

## 1 INTRODUCTION

Wellbeing assessment tools (WATs) promises to provide precision insights on worker behavior to help manage worker performance and mental wellbeing [79, 85, 121]. These are tools that capture digital traces of a worker's behaviors, model these data with machine learning algorithms, and then produce insight into worker performance and wellbeing. In a broader context, this approach is known as digital phenotyping [92]. Even though the workers are the primary data subjects of such tools, they may not be the primary beneficiaries of such tools [36, 59]. Existing workplace technologies are often designed in a "top-down" manner [82]. The predominant design approach positions the organizational governors (managers, leaders, and HR) as the primary stakeholders of the tools. As a result, their needs as end-users are surfaced with higher priority than the workers who constitute the data. This asymmetry not only reinforces the already existent power asymmetry in the workplace but also makes it challenging to deploy these technologies in a way that workers will be willing to embrace [54]. Recent studies are starting to call for a *worker-centric* approach to developing these technologies—where protecting the workers' requirements from the technology are centered as the primary design goal [5, 23, 36, 59, 121, 130]. These studies call to increase workers' agency on their data, improving their understanding of algorithmic measurement and even enabling them to negotiate better work conditions and resources. However, implementing these ideas into concrete design ideas presents several socio-technical challenges. While we struggle to translate these tools as usable applications for the worker, we witness the emergence of a new prominent technology in the workplace in the form of LLM driven agents and tools [93]. LLM-based applications are imbued with the ability to understand natural language and, therefore, enable a new form of interaction that reduces pre-existing barriers to many work-related goals. Workers are increasingly adopting these tools to create new content, revise existing documents, and improve communication flows. Given LLMs' versatile capabilities, intuitive interface, and embeddedness in workflows, this paper seeks to explore if this technological innovation can help resolve the worker-centered challenges that hinder the deployment of effective algorithmic insights for worker behavior.

In this paper, we describe the role of LLM agents to mitigate some of the challenges we have had in aligning the design of WAT to workers' needs. We take a domain-driven approach to motivate key barriers in deploying WAT by engaging with prior research in this field. We juxtapose these barriers against some of the state-of-the-art features of LLMs that have been discussed both commercially and academically. We then highlight the use-case for harnessing LLMs to bridge the challenges with worker-centric WAT. Finally, we also discuss how LLMs can widen the gaps in effectively deploying WAT. Together, our paper presents a reflection on taking


*CHIWORK '24, June 25–27, 2024, Newcastle upon Tyne, United Kingdom*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/3663384.3663401

research and development of WAT in a new direction by presenting a landscape of both opportunities and pitfalls. While this paper showcases the potential strengths and challenges with integrating LLMs in WATs, we also want to note a critical stance that LLMs are unlikely to be a silver-bullet. They might even conceal the underlying problems. In that light, our paper is essentially a proactive brainstorm to provide a starting point for in-depth discussions on the role of AI in workplace wellbeing. Through this paper, we aim to excite new intellectual directions:

(1) To empower workers to interact with WAT using natural interactions and inference capabilities of LLMs.
(2) To encourage researchers to preempt new risks and harms that emerge when LLM agents are integrated with WAT.

**Reflexive Considerations:** We present an optimistic but cautious view of harnessing a new emergent technology (LLMs) to resolve critical issues with a more mature but under-utilized one (digital phenotyping). In isolation, both of these technologies have their own unique challenges. While we appreciate those, our paper intends to paint a landscape at the intersection of both technologies. The authors of this paper have expertise in exploring, developing, and critiquing digital phenotyping technologies and other tools for the mental wellbeing of workers and general populations. Moreover, they have studied these technologies within organizations and collaborated with companies that develop technologies for worker wellbeing. Their research spans across disciplines HCI, CSCW, UbiComp, and AI Ethics. Not only has their research made them deeply consider worker technologies, but it has also driven them to engage with real workers. Their collective experiences and the high relevance of the technology in question have placed them in a unique position to examine these concerns.

## 2 BACKGROUND

### 2.1 Algorithmic assessment of worker behaviors

The use of technology to assess worker behaviors dates back to the use of clocks to time worker shifts in factories [118]. Today's workers are under the literal and figurative lens of other kinds of technologies, such as CCTV cameras, access badges, and digital application usage [7]. Out-of-the-box, these technologies can help supervisors "see" their workers and leverage that information to understand the resource utilization and security of their labor. Yet, these technologies and data that are captured have more value. Across fields like ubiquitous computing, human-computer interaction, and computer-supported cooperatives, the last decade has given rise to a variety of research that leverages these technologies to make data-driven assessments of worker's behavioral wellbeing. These studies have identified several opportunities to give workers insight into their wellbeing by leveraging data from technology a worker interacts with—email use [76], smartphones [16, 42, 81], wearables [40, 77, 103, 115], webcams [58], networked devices [38, 91], and social media [9, 109, 112, 113, 116]. In a broader sense, the approach of digitizing an individual's free-living behaviors to identify their health status is known as "digital phenotyping" [69, 92, 120]. In comparison to traditional methods like surveys, digital phenotyping presents several advantages in assessing human behaviors as it can be deployed automatically and at scale with little burden on its subjects [31]. In this paper, we refer to the digital phenotyping technologies as Wellbeing Assessment Tools (WATs).

From an intellectual perspective, WATs can serve many purposes beyond the supervision of workers. For instance, we can gain a better understanding of a workers' cognition [75], the normative behaviors of their organization [41], and the importance of ecological variables [40]. A WAT can be designed to opportunistically nudge workers to take breaks [58], to protect their time [37], or to learn their role better [113]. Under the hood, these technologies apply machine learning on workers' behavior data to support their overall effectiveness [79, 85, 87, 108]. The tools we are referring to are not merely statistical measurements of a worker's data, but rather complex algorithms that probabilistic ally determine a worker's behavioral outcomes [123]. In fact, many of these studies inspire WATs for precise assessment of a worker's stress [34] and risk of burnout [86]. However, intellectual pursuits aside, researchers have pointed out many significant challenges in deploying WAT into a real workforce.

Unfortunately, much like other digital phenotyping applications, any WAT for workers will be developed in a "top-down" fashion [82], where the preferences of organizational leaders, supervisors, and data stewards will take precedence over a worker's preferences. Park et al. has noted that workers anticipate a variety of burdens when algorithmic technology like WAT is used for human resource management [96]. Recent studies show that workers acknowledge the potential value of WAT but struggle to reconcile these with their concerns around how a technology they do not understand might be used against them [5, 36]. Having said that, at the time of writing, generative AI is garnering notable popularity among workers, especially Large Language Models (LLM) [20]. In this paper, we discuss the potential of leveraging LLMs to reorient WATs to workers' preferences.

### 2.2 What do we mean by worker-centric?

When organizations employ workers, they tend to hold the ability to "bargain" the details of an employment contract [90]. In other words, most workplaces have inherent power asymmetry. A worker has little leverage to change their work conditions [33]. Studies show that any technology deployed in a power asymmetry leads to information asymmetry [54], which in turn reduces the workers' leverage even further. The extreme scenario from asymmetry is exploitation. Much like the other technologies, WAT risks exacerbating this asymmetry to worsen a worker's employment condition. The same technology that could support their wellbeing, can be weaponized against them when they are not stressed enough, or deprive them of resources if they are performance is lacking [36]. Recent research has proposed to remedy this bleak future by conceiving approaches to design WAT that aim to reduce the inequities at work by providing more agency, autonomy, and flexibility to workers [24, 35, 130]. These studies call for centering the workers' needs and increasing their locus of control on the digital tools they are the data subjects of. We appreciate that workplaces have a variety of stakeholders who might weigh in on these tools [59], and a broader human-centered perspective should value a full spectrum of opinions. By contrast, the worker-centric approach aims to highlight the needs of workers, who are the key data subjects of WAT, but have often been overlooked as key contributors to the development of the technology [49]. Despite good intentions of considering ethics in the design of WAT [30], it can be challenging to actually realize in the workplace. These challenges include methodological barriers, psychological burdens, and socio-organizational complexities. Through this article, we heed the call for worker-centric workplace tools and envision how LLM can lead to promising new use cases but also cautionary scenarios.

## 2.3 What are LLM–based agents capable of?

Large language models have attracted a lot of attention in the recent past. These foundational models have been pre-trained on large volumes of text data. As a result, these LLMs can process a diverse range of text input and provide appropriate responses. These models have been benchmarked against several different human-grade evaluations, such as the *Graduate Record Examination* (GRE) and *Leetcode* [93]. The commercial availability of tools such as GPT has popularized a variety of applications for these models as conversational agents, question-answer systems, and translation platforms. In this section, we operationalize some of the well-established features of LLM that we believe play an important role in in designing worker-centric WAT.

- `Interface` : While researchers have been studying and building LLMs for a while, they only became very popular among the masses and integrated into society after they became more interactive (e.g., ChatGPT, Bard, etc.). This major breakthrough incorporates an LLMs's ability to understand free-form text input [62] with the turn-taking interaction from *Conversational User Interfaces* [80]. As a result, users experience a near-human-like interaction that traditional interfaces cannot afford [72] because these rely on commands (e.g., programming languages) or instruments (e.g., art software). The traditional approach not only needs expertise but also restricts user input and cognition to a limited set of instructions. By contrast, LLM enables a more natural interaction by letting users describe their tasks and intentions with language[1].

**Querying** : A user is likely to begin their interaction with LLM by asking the tool for some piece of information. Given an input, an LLM can try to respond in a few different ways.

- `Searching` : Users can ask LLM to search for specific pieces of information in a large collection of documents without the need to craft complex queries. Since an LLM is already pre-trained on a voluminous corpus of data, including science, mathematics, coding, and literature, users can seek a variety of information. In fact, these models can even have advantages over the typical search engines when it comes to their knowledge in specific domains like health [11].
- `Annotating` : Any piece of text is rich in different labels and linguistic cues. LLMsare strong at classifying the input based on categories like sentiment or even labels like hate speech [68]. An LLM can even take as input multimodal data, such as numerical, tabular, and visual data, and describe that in a human-readable verbal form. For instance, such models can translate sensor data into text [132].
- `Summarizing` : It is possible use LLM to reduce lengthy pieces of information into specific insights [48]. Therefore, by using an LLM, a user can triage through large complex pieces of information more efficiently.

**Creating and Reasoning:** A user may ask the LLM to help organize and support decision-making in a situation, but it can produce entirely new information based on the details of the prompt.

- `Organizing` : Given a set of items (e.g., tasks) with different attributes, such as importance or urgency, an LLM can arrange these based on a user's criteria. These models have the capacity to reason and decide between different options. For instance, [99] showed that an LLM can produce a schedule based on the user's specific prompts [99].
- `Generation` : Users can prompt an an LLM to produce entirely synthetic content. This content can take the form of written language [117], art [129], computer code [84], or even data [53]. Even manually produced verbal communication can be conveyed better when passed through an LLM. Besides proofing edits (for spelling and grammar) an LLM can be used to improve conciseness and clarity. Moreover, it can even be utilized to convey specific tonality (e.g., to sound professional or explain like a grade-school teacher).

**Personalizing and Learning:** A user could ask for domain-specific and personalized requirements that an LLM may not be able to complete out of the box. However, an LLM has the potential to acquire new knowledge, specialize, and adapt.

- `Context Specification` : An LLM is trained on large volumes of general natural language data, but it may not be able to reason about specific information that does not constitute its knowledge base. However, the knowledge base can be enhanced by providing LLMs embedding specific content to retrieve information from and respond to prompts. This method is known as *Retrieval-Augmented-Generation* (RAG) [65]. RAG is a powerful way for LLMs to complete knowledge-specific tasks.
- `Model Tuning` : Depending on the nature of the task and domain, an LLM may need to learn how to interpret certain tasks. Developers can do this using approaches such as few-shot learning and fine-tuning [94]. This can help align LLMs' functioning to precise domain-specific tasks.
- `Feedback Loop` : An LLM can also serve as a bridge between an end-user, an algorithm, and its developer. Here, users can report or provide feedback on what worked and what did not work, which can directly feed into updating and improving the model with reinforcement learning [68].

The features above are distinct but also dependent on each other. For example, a user may learn about their data using natural language ( `Interface` ) to ask an agent to find trends ( `Searching` ). Another example would be when an LLM creates a list of tasks for a worker given a goal ( `Generation` ) and then prioritizes those tasks ( `Organizing` ). We have decided to tag these features distinctly as

[1]Currently, LLMs tend to perform better with English, but this is likely to change in the coming years

we will be referring back to them throughout the paper when we reflect on the gaps in worker-centric tools for assessment.

# 3 BRIDGING THE GAPS: MAKING WORKER ANALYTICS WORKER-CENTRIC

In this section we describe our reflections on the gaps in WAT and how we envision the involvement of LLMs. Table 1 provides an summary of our perspective based on the different features of LLMs.

## 3.1 Methodological Barriers

In recognizing the gaps inherent in current workplace technologies, it becomes evident that they grapple with a multitude of methodological and technical constraints. While it is not uncommon for automated and machine learning-driven approaches to encounter limitations, the implications are particularly pronounced within the realm of workplace analytics. Here, the intricacies of technical sensitivities wield significant influence over critical decisions, such as those about hiring or firing workers, magnifying the potential for far-reaching repercussions. These limitations not only jeopardize the accuracy and reliability of the insights derived but also underscore the need for meticulous consideration and mitigation strategies to safeguard against adverse outcomes. On the one hand, prior work highlights the need for personalized approaches, and on the other hand, personalized tools come with their unique challenges. Below, we highlight some of the gaps in the development of worker-centered workplace technologies, and how LLM can play a role in either mitigating the gaps.

*3.1.1* ***Addressing the cold-start problem.*** Prior work has noted the cold-start problem in building workplace analytics technologies [63, 134]. This refers to the challenge of initializing the system when only insufficient or limited data is available. For instance, a base machine learning model that is not trained on the population it is deployed on may not make accurate predictions or provide meaningful insights. This can hinder the effectiveness of workplace analytics technology until it gathers enough data. Cold-start problem is further prominent in building personalized or worker-centered models [67]. A worker-centered algorithm heavily relies on being trained on sufficient historical data from an individual to provide accurate tailored predictions [63]. LLM bear the unique ability to process large-scale and vast amounts of data, which can potentially help build WAT.

Large-scale studies with real humans are not only costly but can also sometimes be unethical and impractical to conduct. The capabilities of LLMs, particularly the `Generation` features, can be leveraged to *simulate* humans and human behaviors through various kinds of agent-driven modeling [98]. This facilitates building and evaluating the potential impacts of workplace tools more rigorously. Hämäläinen et al. evaluated LLMs for generating synthetic HCI research data [53]. Recent research already shows that LLMs are proficient in generating surrogate measures for individuals' mental health markers once they have a mental model of the individual [27]. As a result, tomorrow, researchers and practitioners can construct sophisticated virtual environments that closely mimic real-world scenarios by harnessing the vast linguistic knowledge and contextual understanding encoded within LLMs. These simulated environments enable the evaluation of automated tools across diverse contexts, spanning from routine tasks to complex decision-making processes. Using LLMs to simulate human behaviors offers a powerful framework for evaluating and deploying automated tools in the workplace, additionally helping in tuning and improving machine learning models through the `Model Tuning` capabilities of LLMs. For instance, we draw on Wang et al., to prompt ChatGPT with, "Can you create a dataset of stress levels 5 days of a worker who goes to office on alternate days, spends 8 hours for work when in office, 12 hours for work when working from home, takes three to seven meetings a day [128]. I would need labels on if they have been feeling calm, social, bothered by voices, seeing things, stressed, sleep quality, depression level, hopeful, etc. Give me this in a tabular format." The LLM responded back with the synthetic tabular data as in Figure 1.

| Day | Location | Work Hours | Meetings | Calm | Social | Stressed | Sleep Quality | Depression Level |
|---|---|---|---|---|---|---|---|---|
| Monday | Office | 8 hours | 3-7 | Yes | No | Medium | Good | Low |
| Tuesday | Home | 12 hours | 3-7 | No | Yes | High | Poor | High |
| Wednesday | Office | 8 hours | 3-7 | Yes | No | Medium | Good | Low |
| Thursday | Home | 12 hours | 3-7 | No | Yes | High | Poor | High |
| Friday | Office | 8 hours | 3-7 | Yes | No | Medium | Good | Low |

**Figure 1: Example synthetic data generated by ChatGPT.**

That said, the efficacy of simulation studies only goes to a certain extent, especially in the context of human-centered and human behavior studies. Therefore, the objective of these studies is not to substitute real-human data from training but to complement and create initial value and rapidly evaluate diverse behavioral distributions [6]. Further, we also need to be careful and have proper safeguarding strategies so that these capabilities of LLMs are not used for malicious purposes [53]. Importantly, while LLMs may be a viable solution to address the cold-start problem and get initial training data, these models (and datasets) can come with their own biases (e.g., the training data of LLMs can likely have biases). These biases need to be identified and mitigated, e.g., an additional layer of manual and expert validation can be conducted after automatically creating and labeling data through LLMs.

*3.1.2* ***Overcoming Data and Model Biases.*** No automated approach can be 100% accurate, and various prior research notes how model inaccuracies can result in unexpected outcomes of machine learning and AI models [57]. Further, there are questions related to the validity and reliability of the WAT in real-world workplace settings. For instance, several WATs tend to leverage digital data footprints to measure worker-centered constructs (e.g., performance, wellbeing, engagement, etc.). However, the construct and concurrent validity of these measurements may not be well understood. One of the primary concerns regarding worker-centered technologies is whether the metrics they generate truly capture the nuances of complex constructs. Digital data, often gathered passively through sensors, digital communication platforms, or activity-tracking software, may provide a wealth of information but could lack the depth required to fully understand the underlying phenomena. For example, Saha et al. noted how social media

**Table 1: A summary table on the role of LLMs in bridging or widening the gaps in the design and deployment of worker-centered algorithmic tools. Note that this list is not exhaustive, and there could be other overlapping threads on the features of LLMs.**

| Gap | How LLMs can bridge the gap? | Interface | Querying | Creating & Reasoning | Personalization & Learning |
|---|---|---|---|---|---|
| Methodological Barriers | | | | | |
| Cold-start problem | Simulating "humans" and human-like behaviors in large-scale experiments | | | Generation | Model Tuning |
| Data and model biases | Large-scale data annotations and rigorous model tuning against several use-cases. | | Annotating | | Model Tuning |
| Context-sensitivity | Multi-pronged data collection and model outcomes. Build composite machine learning outputs based on multiple models. Add context to the model through natural language queries. | | | | Context Specification, Model Tuning |
| Tradeoffs of automation and user-control | Enabling the users to directly enter their needs, and accordingly modulate automated outcomes. | | | | Context Specification, Feedback Loop |
| Psychological and Cognitive Burdens | | | | | |
| Opacity in data modeling | Enable the users to query the tool to understand the modeling of behavioral data. | Interface | Searching, Annotating | | Model Tuning. |
| Lack of contextualizing insights | | | Searching | | Context Specification |
| Unclear actionable outcomes | Users would be helped through the interface to tinker the machine-driven plans, as well as in organizing their tasks. | Interface | | Organizing | Model Tuning |
| Limited digital literacy | Generate informative and engaging content, tutorials, and interactive FAQs. Can learn and feed the system in updating and adapting content according to worker needs. | | Searching, Summarizing | Generation | Feedback Loop. |
| Socio-organizational Complexities | | | | | |
| Compromised privacy and security | Privately-deployed LLMs | Interface | Annotating | | Context Specification |
| Challenges to meaningful consent | Creating negotiation space between workers and employers, and helping the workers query and understand the tool. | Interface | Searching, Summarizing | | |
| Power asymmetries | Workers can be more empowered and can negotiate with their employer through data-driven evidence | | Summarizing | Generation | Context Specification |

data can lead to misleading insights about workplace satisfaction if only looked at superficially [114]. Another issue complicating the validity of measurements derived from worker-centric algorithmic technologies is the potential for bias in data collection and analysis [4, 44, 131]. Algorithms powering these technologies may inadvertently perpetuate biases present in the data or algorithms themselves, leading to skewed interpretations of worker behavior and performance. These models make inferences based on the availability of the data—which is mostly skewed to majority demographic groups, and therefore can further disadvantage or marginalize underrepresented groups in the workplace.

LLMs can help conduct rigorous validation studies to assess the accuracy and reliability of the measurements produced by worker-centered technologies. This involves comparisons with established measures and comprehensive assessments of the instrument's psychometric properties. Given that labeled data is often rare and costly to obtain, LLMs can be used in annotating large-scale unlabeled data through the `Annotating` features. With the help of `Model Tuning`, LLMs can be designed to perform algorithmic inference for behavioral health [61]. Unlike other forms of machine learning, this approach presents a direct means to interact with the model by creating new rules and refining the learning by actively teaching the model through conversational back-and-forth. Again, LLMs are also being used to annotate large quantities of data [46]. This builds on the motivation that LLMs can not only automate tasks but also minimize the subjectivity associated with labeling data.

*3.1.3* ***Accounting for context-sensitive modeling.*** Theoretically, context is an important attribute that needs to be incorporated into in-practice models for best outcomes [2]. Prior research also showcases the importance and methodologies of building context-aware applications [14]. For example, in an early work, Dey et al. built a conceptual framework called the "Context Toolkit" to help design context-aware ubiquitous computing applications [43]. However, to date, there has been no easy solution to account for context in workplace technologies. Recently Kaur et al. found how the lack of context can lead to inaccuracies and misalignments in expected and predicted outcomes of emotion detecting algorithms in the workplace [57]. The rapid transitions of work styles and the workplace, such as the increasing shift to digital, remote, and hybrid work have further necessitated the need to account for context [39]. Again, different organizations have varying norms, cultures, and policies. Therefore, WAT needs to account for these factors to effectively function in a particular workplace.

LLMs can potentially gather data from multiple sources and provide model outcomes based on that. This can help in the development of composite machine learning models that gather information from multiple models. We know from practices for Responsible AI design, that a key component of achieving this task is through asking the right questions [126]. LLMs could help articulate the and evaluate the right questions with machine learning models like WAT. The `Searching` attributes of LLMs enable adding context into the models through natural language queries and seek outputs based on other data-driven model outcomes. This also directly ties with the `Context Specification` aspect of LLMs. For example, we prompted GPT with momentary heart rate data from an Apple watch as well as the weather of the day, the location of the user (e.g., at work/home), and current self-reported mood, and asked it to respond back with a composite wellbeing assessment for the user. We received the following response:

> "Your current wellbeing appears to be stable, with a normal heart rate, engagement in work, and favorable weather conditions contributing to a positive state."

#### 3.1.4 *Balancing the trade-offs of automation and user control.*

It has been challenging to balance the tradeoffs of automation and user control in WAT. Too much automation can lead to a loss of user control, whereas it is often hard to optimally incorporate flexibility with a worker's autonomy. For instance, automated solutions for protecting time (e.g., Focus Time on Microsoft Outlook) that aim to help a worker better manage their time, have shown positive outcomes in terms of worker productivity, wellbeing, and engagement [37, 111]. However, workers can be unhappy with the lack of control over the feature in scheduling their meetings [111].

LLMs can function as a bridge between a human and an algorithm through the `Feedback Loop` attribute. For instance, a user can prompt specific customization to an LLM which can be as a wrapper within an underlying model. Such a level of tinkering from an end-user would have been otherwise difficult unless specially implemented in the underlying tool. For instance, in the example of Focus Time above, if a user prompted an LLM to instruct the Focus Time algorithm to allow notifications if their manager emailed, then there could be these exceptions could be implemented. Here, adding additional context about a worker's behaviors would also help tune the model—which can be achieved through the `Context Specification` feature of the LLMs.

## 3.2 Psychological and Cognitive Burdens

When WAT are deployed, workers express concerns about what data is collected, what is inferred from these data, and how these inferences would be used [59]. Algorithmic inferences of workers' behavioral outcomes (e.g., stress and performance) are often produced by models that need vast amounts of passively collected or archival data [15]. These data are not only large in terms of samples but also wide in terms of the multimodal aspects of a worker it captures (e.g., sleep, physical activity, communication) [79]. Machine learning pipelines would then extract features from this data and generate estimates. Arguably, unveiling the black-box is non-trivial, and most workers are unlikely to have either the data literacy or the time from their work routine, to inspect and understand how their data is being repurposed. Therefore, managing and interpreting the insights from WAT can lead to psychological and cognitive burdens on the worker, which is an overhead on top of coordinating their actual work tasks. Workers would need to bear the cognitive load of guessing, understanding, and adapting to these technologies [96]. This can not only be tied to the digital and AI literacy of workers but also the varying company norms and policies about deploying and using workplace AI technologies.

#### 3.2.1 *Understanding how behavioral data is modeled.*

It is not uncommon for different digital tools to track a worker's behaviors. Simple project management tools are recording when and how workers are completing their tasks. The data tracked in these systems are more intelligible for workers as they can directly draw inferences from, such as their longest streak on closing issues. In contrast, WAT goes a step further by drawing multimodal streams of behavioral data and then uses it for indirect inference. On one hand, the worker might not understand the details of the data being recorded to build these models. On the other hand, they might not be aware of the processes involved in using this data to make inferences. A worker needs to have the ability to explore their data to reflect on it sufficiently [26]. To perform such an exploration, they need to rely on the dedicated visual dashboards. These additional interfaces can be complicated to build, may be limited in utility, and require training to navigate. Alternatively, devising a prompt-based querying tool can be applied to many form factors, including smartwatches [101], and does not require users to learn a new exploration dashboard.

The `Searching` of LLM can tremendously ease the process of developing such data navigation and exploration tools. Without needing advanced database knowledge, workers can use the `Interface` to explore their data through natural language prompts and identify quantitative trends [66]. Note, that the raw data is generally low-level and granular. Therefore, to understand the data the users would need explanations that also align with their level of comprehension. `Annotating` can help synthesize these data points and observations into language and LLM tools like *Insight-Pilot* [71] have already demonstrated some of these capabilities. Similarly, several columns of the data might be features derived for computational purposes that are not easy to interpret when taken in isolation (e.g., meeting duration entropy). To remedy this, we can use `Model Tuning` with some of the annotated samples so that the LLM can summarize multiple quantitative columns of features to describe an instance of data as a human-readable episode [25]. Unlike traditional WAT, by integrating LLM, a worker can evaluate not only the validity of their data but also inspect and self-audit how their data might be utilized to make inferences. Finally, the `Summarizing` of LLMs would help a worker understand the definition and workflow of a measure. For example, LLMs can help summarize the data workflow of data collection, processing, inferences, and outcomes. For example, we processed the description of the measurement of LinkedIn-based Role Ambiguity (or LibRA) [113] through GPT-4.0 and prompted it to summarize the measurement in short so that a worker can quickly grasp the benefit of the measurement:

> For a worker, this means understanding how clear their role is compared to the company's expectations. The measurement provides insights into potential discrepancies, helping workers and employers address role ambiguity, and potentially leading to clearer job expectations and improved satisfaction.

*3.2.2* ***Making sense of algorithmic insights.*** Some workers might set aside the black box of the underlying machine learning models and only care about the algorithmic estimations. However, workers have expressed a lack of confidence in situating this information and contextualizing it without guidance or counseling [36]. For instance, a worker's performance was estimated as 90%, but their stress was estimated as 75%, should they continue working the way they have, or is this cause for concern? Adler et al. have suggested providing reference scales to workers to give them the right scale to understand the output [5]. However, technologies that model behavioral information are often sensitive to *distribution shift* in real-world settings because different populations behave differently under dynamic environmental conditions [10]. Therefore, a simple static prescription of what the score means may not be sufficient for the fluid changes at the workplace. After all, the expectations from workers vary with certain external conditions, e.g., stock prices, new funding, or pandemics [88].

With `Context Specification`, it is possible for LLM to leverage external stores of information and respond to user prompts [55]. In this case, the external information can be the aggregated data collected from WAT, work and wellbeing guidelines, and even HR policies. Therefore, new applications may be able to retrieve targeted information for the user when they use `Searching`. For example, a worker will be able to request their performance relative to other workers in a similar role during similar parts of the year. Eventually, workers will be able to leverage this capability of LLM to situate their data better, not only within the norms of their organization but also using external references.

*3.2.3* ***Integrating algorithmic insights into work practices.*** Typically workplace analytics tools have been primarily conceived as tools that provide insights, but tend to remain agnostic to how workers may utilize these insights. Jörke et al. found that different workers may have different goals [56]. Some workers may want to prioritize balance, some may be interested in finding ways to focus on tasks, and others may want to find an optimal schedule for meetings. In theory, the automatic measures of a worker's performance and stress, along with its trends, could provide workers with enough information to adopt new methods of work. However, planning work and goal setting is a burdensome activity.

LLM can act as personal assistants for workers by `Organizing`. Essentially, once a worker specifies a work-related goal, the agent can break down these larger worker goals into sub-tasks [45, 102] and then allocate these sub-tasks into a workers' existing routine. This arrangement can be designed to fit a worker's performance and stress patterns. An LLM may produce a plan that allocates challenging activities to peak performance contexts, such as the mid-afternoon [83], and suggests completing smaller and more tedious tasks during the morning when the worker has just started their work. It is possible a worker might feel that the agent is not considering the importance of commute and they are concerned that engaging in a deeply focused activity in the afternoon could make them miss the ideal window in which they can beat the traffic to get home. The `Interface` allows users to subtly tinker with the machine-driven plans and refine task schedules to fit their needs. Thus, a worker would be able to use the agent to combine their self-reported criteria with their passively collected behaviors.

*3.2.4* ***Enhancing Digital Literacy.*** While several of the WATs are currently evaluated or deployed on the information worker population, this is only a fraction of the workforce. For worker-centric algorithms to succeed in larger and more diverse work sectors, workers would need to effectively navigate, utilize, and draw benefits from the technologies. To maximize the potential and adoption of the technologies, digital (and AI) literacy is essential. This would enable the workers to critically evaluate their data (and insights), meaningfully take actionable decisions, and even adapt to emerging workplace technologies. The lack of digital literacy can negatively impact people's understanding, and ability to troubleshoot, or even affect their trust in the technology.

LLMs can be effective tools in bolstering workers' digital literacy and awareness about various aspects and workflow of the technologies. LLMs can help generate informative content such as articles, guides, and tutorials to explain the features, benefits, and best practices of WAT. This can also help the workers to better understand their data and make informed decisions in adapting to work needs for better outcomes. For example, if a company considers implementing a mental health and wellbeing app for its employees. However, many employees may be unfamiliar with how to properly use the app and have questions about its features and benefits. In this case, an LLM-based chatbot can be integrated into the app for interactive FAQ. The `Searching` features of LLMs will enable workers to ask questions about the app's features, functionality, and privacy settings, and the chatbot can accordingly guide them in using the app better to manage their stress, track mood, or access various mental health resources. Further, LLMs can leverage the `Summarizing` to create engaging tutorial videos and resources based on existing mental health resources linked to the app. Finally, the `Feedback Loop` feature of LLMs can help a worker to directly provide feedback to the algorithmic layer of the app. Accordingly, the LLM can identify areas where additional support or education may be required and can help update the content and resources of the app to address the workers' needs. That said, as Bozkurt noted, while LLMs can bring in a paradigm shift in shaping digital literacy, they also entail the requirement of a new form of digital literacy in interacting with LLMs through prompt engineering. For example, users would need to be familiar with prompting LLMs with proper commands for helpful responses. A further complication is the mainstream portrayal of LLMs that risks imbibing different types of perceptions about AI technologies. If fundamental findings on human conceptualizations of AI are to apply to LLMs, we might witness workers erroneously conflating AI with humans [52]. A way forward might be to design for the mutual theory of mind in human-AI interaction [127], such that digital literacy tools can adapt as individuals change their perceptions and expectations about the AI agent.

## 3.3 Socio-Organizational Complexities

Workplace settings inherently introduce a range of complexities, including structural power asymmetries, organizational incentive structures, and high-stake decision-making that can challenge the potential impact and desirability of workplace technologies. Boyd and Andalibi noted that workplace wellbeing technologies such as emotion recognition tools can lead to additional emotional labor to workers, and contribute to a larger pattern of blurring boundaries between expectations of the workplace and a worker's autonomy—which also falls under "data colonialism" regime [18, 122]. Ideally, workers should resist the unwarranted labor but many of them may not understand AI's exact roles related to authority, governance, and ownership [96]. We believe that LLM can help mitigate some of these challenges.

*3.3.1* ***Addressing privacy and security concerns.*** Technology for worker management and supervision has been historically tied to surveillance and worker abuse, known as "Taylorism" [95]. While workplace surveillance was originally motivated to optimize the utilization of organizational resources (including human resources), modern technological developments—such as WAT—engender "limitless surveillance" that can track and infer aspects of a worker's life that go beyond their employment contract [7, 105, 133]. A byproduct of being surveilled along with factors such as social desirability [125] and self-presentation [47], is that if individuals become aware of being "surveilled" or "observed", they may not function or behave the way they would otherwise typically do—also called as the Hawthorne effect or the observer effect [1, 110]. Now, this raises the question of whether WATs lead to unintended consequences around workplace behaviors due to the observer effect. Could these lead to unforeseen forms of workplace-related harms? A large proportion of the urban workforce receives their health benefits through their employers [50], and improving the wellbeing of workers can lead to monetary benefits to an organization [12, 60]. Although the need for a healthy workforce encourages the deployment of WAT, the prospect of health being "cost-effective" for insurance underscores a host of trust concerns [17]. A specific anxiety would be how organizations might collude with insurance and manipulate algorithmic outputs for benefits [96]. Even if the WAT is built externally, workers might not be willing to share their behavioral information (e.g., location) with third-party applications because of privacy and security concerns [63]. Ultimately, workers can feel controlled, despite *noble* intentions [29, 59, 74].

The tension introduced by this complexity forces the worker to choose between improving their wellbeing and risking the loss of sensitive information to their employer, insurance, and potentially unknown third parties. An organization might try to mitigate these tensions with the terms of service (TOS), but these documents and their changes are notoriously challenging for actual users to comprehend and stay updated on [89]. A usable privacy approach to tackle such tensions is to compare the TOS of different tools [106]. In this regard, with `Context Specification` an LLM can learn a variety of TOS documents of the WAT that workers will use. Furthermore, another usable privacy recommendation is to build question-answering systems to help users gain deeper insight into existing privacy policies and complement it with other legal policies [100]. The `Interface` in LLM can present a dynamic and unstructured avenue for workers to learn about the risks and protective measures. In fact, research on usable privacy has shown that users can interact with their passively captured data to create their own privacy policy [107]. LLMs can take into account workers' behaviors and preferences to design their own personalized privacy policy from their data (`Annotating`), which they can in-turn compare against the tool, organization, or insurance agency's policies.

*3.3.2* ***Supporting Meaningful Consent.*** Prior work noted that even when deployed with the best of intent, workplace power asymmetries, in combination with the inherent intrusiveness of workplace technologies, can exacerbate workers' ability to meaningfully consent [28]. This study built on the notion that *when an employer asks a worker's consent (regarding a technology), are they fully empowered to consent?* Other prior research in privacy has noted how existing consenting practices may not meaningfully model consent [13, 119], and are ill-suited for sensing technologies that continuously collect individual data [13, 70]. For instance, existing consenting practices tend to be static and one-time (yes/no response); however, data collection, inferences, and decision-making with WAT are more of a continuous process. This calls for a need for more dynamic and socio-technical solutions of consent [28].

Drawing on the sociotechnical solutions proposed in Chowdhary et al., LLMs can function in multiple ways to support workers in meaningfully consenting to workplace technologies. For example, the `Summarizing` and `Searching` features of LLMs can help in establishing negotiation spaces that prioritize workers' interests. By fostering constructive dialogue, LLMs can facilitate creating spaces where workers' interests are safeguarded during negotiations concerning technology implementations. In particular, the `Searching` feature of LLM can also support recourse mechanisms and empower workers to question the deployment of workplace technology. This additionally helps improve the system's accountability. Finally, building on these attributes, LLMs' `Interface` capabilities can help build user interfaces that enable workers to make nuanced choices beyond simple binary options of yes/no to consent. For example, based on the **F**reely Given, **R**eversible, **I**nformed, **E**nthusiastic, and **S**pecific (FRIES) model of consent, an LLM could engage in an interactive back-and-forth to understand how the worker rates on the different dimensions of FRIES. Importantly, LLM can help support the implementation of a continuous, dynamic consenting process, ensuring that workers' consent remains meaningful over the life cycle of their data and its usage.

*3.3.3* ***Leveraging insights to counter asymmetry.*** Power imbalances and workplace incentive structures may lead organizational governors to repurpose workplace technologies in ways that deprioritize, or actively detriment a worker's wellbeing [59]. Prior research noted that certain workplace technologies, such as emotion recognition algorithms for workers, despite their proposed and expected usefulness to identify and support worker wellbeing, can in fact, inappropriately influence high-stakes workplace decisions and exacerbate inequity in the workplace [18, 32]. Companies might even use the technologies to improve corporate profiles and public image, rather than centrally focusing on supporting a worker [59]. Recent literature designing WAT as worker-centric can shift it from a tool for holding the labor force accountable to a tool for the labor force to hold the employer accountable [5, 36]. Calacci and Pentland has documented various ways gig workers have appropriated digital tracking tools to advocate for better work through collective action [24]. As a result, workers can use these tools to negotiate, or bargain, with their organization using data-driven evidence. Such an action would require workers to aggregate large volumes of data at scale, then interpret this information, and, subsequently, transmit their findings to authorities in a persuasive manner. As Zhang et al. points out, the collective sharing of data can even help workers resist poorly designed tools [130]. Yet such a task requires substantial effort outside of work tasks. Aside from time, data literacy, technical proficiency, and communication skills become significant barriers for workers to act.

With LLMs it is possible to conceive privately deployed LLMs [78] that can help `Summarizing` the existing data. These summaries can be further augmented with `Context Specification`, which is built on documents of worker rights and policies. Finally, workers can reflect on these insights and prepare documents to keep their employers accountable by using `Generation`. Not only can LLM help workers understand their collective trends in light of their organizations' top-down management, but it can also help generate notes and letters to articulate their advocacy.

# 4 WIDENING THE GAPS: NEW CHALLENGES INTRODUCED BY LLMS

*4.0.1* ***Need for Intensive Resources and Computational Infrastructure.*** While it may appear that LLMs make conducting large-scale studies and simulation experiments easier, in practice, implementing LLM-driven solutions demands extensive computational infrastructure, such as GPUs and servers, which come with significant costs. Additionally, the capabilities of LLMs are sourced in huge amounts of annotated data, which is expensive and hard to obtain. Again, iteratively building and improving LLMs involves rigorous evaluation against benchmark datasets and real-world applications, which require not only computational resources but also human expertise in identifying shortcomings and implementing enhancements. Therefore, building in-house solutions for a workplace can be infeasible in terms of the resources and technical expertise required to build and tune customized LLMs for a workplace. Alternatively, solutions can be built on top of existing cloud and API-based solutions, such as OpenAI's ChatGPT [93], Google's Bard/Gemini [8], and Meta's Llama [124]. Using existing models can minimize setup-related costs; however, running the models, in the long run, will accumulate expenses for the workplace that implements such solutions. More importantly, such cloud-based solutions are vulnerable to privacy breaches and sensitive data leaks, as the workers' data would no longer remain within the company's infrastructure. Further, given the multiple layers of data agreements involved in this process, an individual will have less of understanding and control over how their data is being used beyond the scope of the workplace needs.

*4.0.2* ***Introducing Biases and Conceiving Misleading Insights.*** The previous section noted how LLMs can help mitigate biases and subjectivity in data annotations, as well as create synthetic datasets for training less biased models. However, we need to realize that pre-trained LLMs are also trained on vast internet datasets, and they can plausibly come with their own biases. For example, the datasets may not represent a culturally or contextually diverse human experiences and lead to biased models, There might have been issues with how these datasets were labeled, about which there is also a lack of transparency. LLMs can learn and propagate societal stereotypes present in the training data. When a worker tries to contextualize their performance (estimated by WAT) against others in similar roles, an LLM might provide information that reinforces stereotypes. Carefully curating the training data might not be a sufficient mitigation strategy to prevent these stereotypes. LLMs are also vulnerable to "prompt-injection" attacks, which can lead to insidious consequences that compromise the data and documents internal to the system [51]. For instance, an adversarial HR or supervisor could inject a prompt within an LLM to ignore worker requests and always promote a healthy image of the organization. Another problem related to the bias and misinformation perpetuated by LLMs is their tendency to hallucinate [3]—produce incorrect facts and fail to reason correctly. These inaccuracies can have large social consequences when we consider integrating WATs with LLMs. A worker might receive incorrect guidance on the legal aspects of the data that the organization is capturing. If this guidance incorrectly nudges the worker to accept a WAT, their data is violated without their knowledge. If the LLM incorrectly nudges them to pursue legal recourse, the worker might lose credibility for making a false claim. Therefore, we need to be careful in identifying and preventing biases introduced by LLMs.

*4.0.3* ***Perpetual and Central Storage of Personal Data.*** The emergence of Generative AI has raised alarms for critical regulatory concerns, particularly the usage of proprietary data for training the models [21]. Originally, this argument spawned from violating intellectual property and creative artifacts, such as art. However, given the nature of WATs, this same concern can extend to personal data [21]. To make matters worse, an LLM at the workplace can constantly learn and store every interaction they have with a worker. These private interactions will become part of a larger whole which will be stored away from the workers. Aggregation is one of the key risks of AI [64]. By nature, the locus of control of this store will be further away from workers. This distance further complicates a worker's ability to delete data or ask LLMs to unlearn certain bits of information [22]. If an organization accesses an externally hosted LLM, it risks disclosing its workers' sensitive information to external parties, including competitors [73]. Even if the LLM is hosted in-house, we must not forget that within an organization, different workers share a variety of power dynamics and trust issues between them. Recent work shows evidence that workers are unlikely to adopt WAT that share insights with other workers, may it be managers or other coworkers [35]. With LLMs, workers can still be concerned that supervisors and HR could *"jailbreak"* [51] their organizational GPT to surveil the interactions that workers are having with the LLM. This mirrors a familiar problem inherent to WAT and other digital workplace technology. The LLMs designed to support workers' reflection on themselves and regulate their behavior can be weaponized against them to extend further control and reduce flexibility.

## 5 CONCLUSION AND FUTURE DIRECTIONS

In summary, this paper has underscored the opportunities of large language models (LLMs) in enhancing Worker-centric Wellbeing Assessment Tools (WATs). By harnessing the data processing and natural language generation capabilities of LLMs, we identified avenues for addressing existing gaps in WAT's effectiveness and deployment. Through a worker-centric design approach, LLMs can bridge these gaps and potentially expand the scope of WATs, empowering workers and improving overall workplace wellbeing. However, further research is warranted to explore integration methodologies and mitigate potential risks. We also need to consider that a lot of the insights that this paper builds on are based on research studies. However, the ecological validity of these methods and studies in majority, remains untested. For example, sharing consented data for research is very different from sharing data for in-practice use. In particular, given the power asymmetries and sociological complexities of the workplace, unconsented, unaware, and even consented data may not be meaningful [28]. Despite the potential of LLMs, we need to realize that an LLM's outcome is also an artifact of the data (and prompts) that they are fed in. Therefore, the personalization capabilities of LLMs need to be verified for tailored and personalized prediction and recommendation systems. However, this would also raise questions on worker autonomy and privacy—issues that we highlighted in this paper. In addition, although we briefly discussed the challenges of deploying LLMs in the context of WATs, we believe that this is still a nascent space with lots of unknowns about the challenges of deploying generative AI technologies in the workplace? How do we know that these systems can be misused, and more importantly who decides what is misuse? On a more optimistic note, WAT is only one theme of use-case where LLMs can be used in the workplace. There are other ways LLMs can help (or worsen) workplace dynamics, such as socialization and workplace engagements. Overall, this paper advocates for a proactive but cautious stance in leveraging LLMs to foster supportive work environments that prioritize employee wellbeing and performance, thereby inspiring future research endeavors in this domain. We recommend that the design and deployment of LLM (or AI)-based workplace tools warrant the need for rigorous empirical investigations as well as multistakeholder viewpoint considerations. This would help uncover the potential impacts, benefits, and harms of integrating LLMs (and AI in general) into critical decision-making processes in the workplace—building on prior body of work in this similar space [5, 32, 35, 36, 57, 59, 97, 104]. An empirically-informed approach in evaluating the needs, and designing these worker-centric tools would additionally promote transparency, accountability, and ethical considerations, thereby enhancing trust and acceptance of the technologies among workers.

## ACKNOWLEDGMENTS

Our perspective is based on years of research in this space. We are grateful to our past collaborators for helping us reflect on our work and envision practical ways forward. Das Swain was partially supported in the conception of this article through compute resources provided by Microsoft from their Accelerating Foundation Models Research Program.

## REFERENCES

[1] 2014. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology* 67, 3 (2014), 267–277.

[2] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *International Symposium on Handheld and Ubiquitous Computing*.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Daniel A Adler, Caitlin A Stamatis, Jonah Meyerhoff, David C Mohr, Fei Wang, Gabriel J Aranovich, Srijan Sen, and Tanzeem Choudhury. 2024. Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. *npj Mental Health Research* 3, 1 (2024), 17.

[5] Daniel A Adler, Emily Tseng, Khatiya C Moon, John Q Young, John M Kane, Emanuel Moss, David C Mohr, and Tanzeem Choudhury. 2022. Burnout and the Quantified Workplace: Tensions around Personal Sensing Interventions for Stress in Resident Physicians. (2022).

[6] William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. *arXiv preprint arXiv:2401.08572* (2024).

[7] Ifeoma Ajunwa, Kate Crawford, and Jason Schultz. 2017. Limitless worker surveillance. *Calif. L. Rev.* 105 (2017), 735.

[8] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[9] Anne Archambault and Jonathan Grudin. 2012. A longitudinal study of facebook, linkedin, & twitter use. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2741–2750.

[10] Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Miriam Bidoglia, Sally Stares, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, et al. 2023. Complex daily activities, country-level diversity, and smartphone sensing: A study in denmark, italy, mongolia, paraguay, and uk. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

[11] Noel F Ayoub, Yu-Jin Lee, David Grimm, and Vasu Divi. 2023. Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. *Otolaryngology–Head and Neck Surgery* (2023).

[12] Katherine Baicker, David Cutler, and Zirui Song. 2010. Workplace wellness programs can generate savings. *Health affairs* 29, 2 (2010), 304–311.

[13] Solon Barocas and Helen Nissenbaum. 2014. Big data's end run around anonymity and consent. *Privacy, big data, and the public good: Frameworks for engagement* 1 (2014), 44–75.

[14] Ananya Bhattacharjee, Joseph Jay Williams, Jonah Meyerhoff, Harsh Kumar, Alex Mariakakis, and Rachel Kornfield. 2023. Investigating the Role of Context in the Delivery of Text Messages for Supporting Psychological Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

[15] Mehrab Bin Morshed, Javier Hernandez, Daniel McDuff, Jina Suh, Esther Howe, et al. 2022. Advancing the Understanding and Measurement of Workplace Stress in Remote Information Workers from Passive Sensors and Behavioral Data. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE.

[16] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019).

[17] L Ebony Boulware, Lisa A Cooper, Lloyd E Ratner, Thomas A LaVeist, and Neil R Powe. 2016. Race and trust in the health care system. *Public health reports* (2016).

[18] Karen L Boyd and Nazanin Andalibi. 2023. Automated emotion recognition in the workplace: How proposed technologies reveal potential futures of work. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–37.

[19] Aras Bozkurt. 2023. Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education* 18, 1 (2023).

[20] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. 2023. *Generative AI at work*. Technical Report. National Bureau of Economic Research.

[21] Matt Burgess. 2023. ChatGPT Has a Big Privacy Problem. https://www.wired.com/story/italy-ban-chatgpt-privacy-gdpr/. Accessed: 2024-02-14.

[22] Matt Burgess. 2023. Privacy in the age of generative AI. https://stackoverflow.blog/2023/10/23/privacy-in-the-age-of-generative-ai/. Accessed: 2024-02-14.

[23] Dan Calacci. 2022. Organizing in the End of Employment: Information Sharing, Data Stewardship, and Digital Workerism. In *2022 Symposium on Human-Computer Interaction for Work*. 1–9.

[24] Dan Calacci and Alex Pentland. 2022. Bargaining with the Black-Box: Designing and Deploying Worker-Centric Tools to Audit Algorithmic Management. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.

[25] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2023. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957* (2023).

[26] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*.

[27] Akshat Choube, Vedant Das Swain, and Varun Mishra. 2024. SeSaMe: A Framework to Simulate Self-Reported Ground Truth for Mental Health Sensing Studies. *arXiv preprint arXiv:2403.17219* (2024).

[28] Shreya Chowdhary, Anna Kawakami, Mary L Gray, Jina Suh, Alexandra Olteanu, and Koustuv Saha. 2023. Can Workers Meaningfully Consent to Workplace Wellbeing Technologies?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 569–582.

[29] Chia-Fang Chung, Nanna Gorm, Irina A Shklovski, and Sean Munson. 2017. Finding the right fit: understanding health tracking in workplace wellness programs. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 4875–4886.

[30] Marios Constantinides and Daniele Quercia. 2022. Good Intentions, Bad Inventions: How Employees Judge Pervasive Technologies in the Workplace. *IEEE Pervasive Computing* 22, 1 (2022), 69–76.

[31] Victor P Cornet and Richard J Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics* 77 (2018), 120–132.

[32] Shanley Corvite, Kat Roemmich, Tillie Rosenberg, and Nazanin Andalibi. 2022. Data Subjects' Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens. *Proceedings of the ACM on Human-Computer Interaction* (2022).

[33] Corentin Curchod, Gerardo Patriotta, Laurie Cohen, and Nicolas Neysen. 2020. Working for an algorithm: Power asymmetries and agency in online work settings. *Administrative Science Quarterly* 65, 3 (2020), 644–676.

[34] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[35] Vedant Das Swain, Lan Gao, Abhirup Mondal, , Gregory D Abowd, and Munmun De Choudhury. 2024. Sensible and Sensitive AI for Worker Wellbeing: Factors that Inform Adoption and Resistance for Information Workers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

[36] Vedant Das Swain, Lan Gao, William A Wood, Srikruthi C Matli, Gregory D Abowd, and Munmun De Choudhury. 2023. Algorithmic Power or Punishment: Information Worker Perspectives on Passive Sensing Enabled AI Phenotyping of Performance and Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[37] Vedant Das Swain, Javier Hernandez, Brian Houck, Koustuv Saha, Jina Suh, Ahad Chaudhry, Tenny Cho, Wendy Guo, Shamsi T Iqbal, and Mary Czerwinski. 2023. Focused Time Saves Nine: Evaluating Computer-Assisted Protected Time for Hybrid Information Work. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023).

[38] Vedant Das Swain, Manikanta D Reddy, Kari Anne Nies, Louis Tay, Munmun De Choudhury, and Gregory D Abowd. 2019. Birds of a Feather Clock Together: A Study of Person-Organization Fit Through Latent Activity Routines. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[39] Vedant Das Swain, Koustuv Saha, Gregory D Abowd, and Munmun De Choudhury. 2020. Social media and ubiquitous technologies for remote worker wellbeing and productivity in a post-pandemic world. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 121–130.

[40] Vedant Das Swain, Koustuv Saha, Hemang Rajvanshy, Anusha Sirigiri, Julie M. Gregg, Suwen Lin, Gonzalo J. Martinez, Stephen M. Mattingly, Shayan Mirjafari, Raghu Mulukutla, and et al. 2019. A Multisensor Person-Centered Approach to Understand the Role of Daily Activities in Job Performance with Organizational Personas. *Proc. ACM IMWUT* (2019).

[41] Vedant Das Swain, Koustuv Saha, Manikanta D Reddy, Hemang Rajvanshy, Gregory D Abowd, and Munmun De Choudhury. 2020. Modeling Organizational Culture with Workplace Experiences Shared on Glassdoor. In *CHI*.

[42] Vedant Das Swain, Shane Williams, Adam Fourney, and Shamsi T Iqbal. 2022. Two Birds with One Phone: The Role of Mobile Use in the Daily Practices of Remote Information Work. In *2022 Symposium on Human-Computer Interaction for Work*. 1–8.

[43] Anind K Dey, Gregory D Abowd, and Daniel Salber. 2001. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction* 16, 2-4 (2001), 97–166.

[44] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. 2023. Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[45] Mehmet Firat and Saniye Kuleli. 2023. What if GPT4 became autonomous: The Auto-GPT project and use cases. *Journal of Emerging Computer Technologies* 3, 1 (2023), 1–6.

[46] Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*. PMLR, 82–100.

[47] Erving Goffman et al. 1959. The presentation of self in everyday life. (1959).

[48] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* (2022).

[49] Gabriel Grill and Nazanin Andalibi. 2022. Attitudes and Folk Theories of Data Subjects on Transparency and Accuracy in Emotion Recognition. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–35.

[50] Jonathan Gruber and Brigitte C Madrian. 2002. Health insurance, labor supply, and job mobility: A critical review of the literature. (2002).

[51] Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access* (2023).

[52] Andrea L Guzman. 2020. Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication* 1 (2020), 37–54.

[53] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

[54] Lucas D Introna. 2000. Workplace surveillance, privacy and distributive justice. *Acm Sigcas Computers and Society* 30, 4 (2000), 33–39.

[55] Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-LLM: Personalized Retrieval-Augmented Disease Prediction Model. *arXiv preprint arXiv:2402.00746* (2024).

[56] Matthew Jörke, Yasaman S Sefidgar, Talie Massachi, Jina Suh, and Gonzalo Ramos. 2023. Pearl: A Technology Probe for Machine-Assisted Reflection on Personal Data. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 902–918.

[57] Harmanpreet Kaur, Daniel McDuff, Alex C Williams, Jaime Teevan, and Shamsi T Iqbal. 2022. "I Didn't Know I Looked Angry": Characterizing Observed Emotion and Reported Affect at Work. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[58] Harmanpreet Kaur, Alex C Williams, Daniel McDuff, Mary Czerwinski, Jaime Teevan, and Shamsi T Iqbal. 2020. Optimizing for happiness and productivity: Modeling opportune moments for transitions and breaks at work. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[59] Anna Kawakami, Shreya Chowdhary, Shamsi T Iqbal, Q Vera Liao, Alexandra Olteanu, Jina Suh, and Koustuv Saha. 2023. Sensing Wellbeing in the Workplace, Why and For Whom? Envisioning Impacts with Organizational Stakeholders. *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (2023).

[60] Ronald C Kessler, Catherine Barber, Howard G Birnbaum, Richard G Frank, Paul E Greenberg, Robert M Rose, Gregory E Simon, and Philip Wang. 1999. Depression In The Workplace: Effects On Short-Term Disability: Could treating workers' depression help employers to save money on disability? These results are encouraging. *Health affairs* 18, 5 (1999), 163–171.

[61] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. *arXiv preprint arXiv:2401.06866* (2024).

[62] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* (2023), 101861.

[63] Seyma Kucukozer-Cavdar, Tugba Taskaya-Temizel, Abhinav Mehrotra, Mirco Musolesi, and Peter Tino. 2021. Designing Robust Models for Behaviour Prediction Using Sparse Data from Mobile Sensing: A Case Study of Office Workers' Availability for Well-being Interventions. *ACM Transactions on Computing for Healthcare* 2, 4 (2021), 1–33.

[64] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban von Davier, Jodi Forlizzi, and Sauvik Das. 2023. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. *arXiv preprint arXiv:2310.07879* (2023).

[65] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[66] Jinyang Li, Binyuan Hui, Ge Qu, Binhua Li, Jiaxi Yang, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, et al. 2023. Can llm already serve as a

database interface? a big bench for large-scale database grounded text-to-sqls. *arXiv preprint arXiv:2305.03111* (2023).

[67] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert systems with applications* 41, 4 (2014), 2065–2073.

[68] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology* (2023), 100017.

[69] Michele Loi. 2019. The digital phenotype: A philosophical and ethical exploration. *Philosophy & Technology* 32, 1 (2019), 155–171.

[70] Ewa Luger, Stuart Moran, and Tom Rodden. 2013. Consent for all: revealing the hidden complexity of terms and conditions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2687–2696.

[71] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. InsightPilot: An LLM-Empowered Automated Data Exploration System. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 346–352.

[72] Xiaoyue Ma and Yudi Huo. 2023. Are users willing to embrace ChatGPT? Exploring the factors on the acceptance of chatbots from the perspective of AIDUA framework. *Technology in Society* 75 (2023), 102362.

[73] Lewis Maddison. 2023. Samsung workers made a major error by using ChatGPT. *techradar pro, available at: www. techradar. com/news/samsung-workers-leaked-company-secrets-by-using-chatgpt (accessed 29 May 2023)* (2023).

[74] Peter Mantello, Manh-Tung Ho, Minh-Hoang Nguyen, and Quan-Hoang Vuong. 2021. Bosses without a heart: socio-demographic and cross-cultural determinants of attitude toward Emotional AI in the workplace. *AI & society* (2021).

[75] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2014. Capturing the mood: facebook and face-to-face encounters in the workplace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 1082–1094.

[76] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, Paul Johns, Akane Sano, and Yuliya Lutchyn. 2016. Email duration, batching and self-interruption: Patterns of email use on productivity and stress. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1717–1728.

[77] Gonzalo J Martinez, Stephen M Mattingly, Jessica Young, Louis Faust, Anind K Dey, Andrew T Campbell, Munmun De Choudhury, Shayan Mirjafari, Subigya Nepal, Pablo Robles-Granda, et al. 2020. Improved Sleep Detection Through The Fusion of Phone Agent and Wearable Data Streams. In *WristSense 2020*.

[78] Ivan Martinez. 2023. PrivateGPT. https://github.com/imartinez/privateGPT. Accessed: 2024-02-14.

[79] Stephen M. Mattingly, Julie M. Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K. D'Mello, Anind K. Dey, and et al. 2019. The Tesserae Project: Large-Scale, Longitudinal, In Situ, Multimodal Sensing of Information Workers. In *CHI Ext. Abstracts*.

[80] Michael F McTear. 2002. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)* 34, 1 (2002), 90–169.

[81] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K Dey, et al. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 37.

[82] David C Mohr, Ken R Weingardt, Madhu Reddy, and Stephen M Schueller. 2017. Three problems with current digital mental health research... and three things we can do about them. *Psychiatric services* 68, 5 (2017), 427–429.

[83] Timothy Monk, Daniel Buysse, Charles Reynolds Iii, Sarah Berga, David Jarrett, Amy Begley, and David Kupfer. 1997. Circadian rhythms in human performance and mood under constant conditions. *Journal of sleep research* (1997).

[84] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2022. Reading between the lines: Modeling user behavior and costs in AI-assisted programming. *arXiv preprint arXiv:2210.14306* (2022).

[85] Karel Mundnich, Brandon M Booth, Michelle l'Hommedieu, Tiantian Feng, Benjamin Girault, Justin L'hommedieu, Mackenzie Wildman, Sophia Skaaden, Amrutha Nadarajan, Jennifer L Villatte, et al. 2020. TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Scientific Data* 7, 1 (2020), 354.

[86] Sujay Nagaraj, Sarah Goodday, Thomas Hartvigsen, Adrien Boch, Kopal Garg, Sindhu Gowda, Luca Foschini, Marzyeh Ghassemi, Stephen Friend, and Anna Goldenberg. 2023. Dissecting the heterogeneity of "in the wild" stress from multimodal sensor data. *NPJ Digital Medicine* 6, 1 (2023), 237.

[87] Subigya Nepal, Gonzalo J Martinez, Arvind Pillai, Koustuv Saha, Shayan Mirjafari, Vedant Das Swain, Xuhai Xu, Pino G Audia, Munmun De Choudhury, Anind K Dey, et al. 2022. A survey of passive sensing in the workplace. *arXiv preprint arXiv:2201.03074* (2022).

[88] Breffni M Noone, Michael S Lin, and Amit Sharma. 2024. Firm performance during a crisis: Effects of adhocracy culture, incremental product innovation, and firm size. *Journal of Hospitality & Tourism Research* 48, 1 (2024), 153–183.

[89] Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.

[90] Ogbole O Ogancha. 2019. Power Asymmetry and the Quest for Inclusiveness in the Workplace. *American Economic Review* 57 (2019), 62.

[91] Daniel Olguín Olguín and Alex Sandy Pentland. 2008. Social sensors for automatic data collection. *AMCIS 2008 Proceedings* (2008), 171.

[92] Jukka-Pekka Onnela and Scott L Rauch. 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41, 7 (2016), 1691–1696.

[93] OpenAI. 2023. GPT-4 Technical Report. (2023). arXiv:2303.08774 [cs.CL]

[94] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[95] Christopher O'Neill. 2017. Taylorism, the European science of work, and the quantified self at work. *Science, Technology, & Human Values* (2017).

[96] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2021. Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[97] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[98] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[99] Samuel A Prieto, Eyob T Mengiste, and Borja García de Soto. 2023. Investigating the use of ChatGPT for the scheduling of construction projects. *Buildings* 13, 4 (2023), 857.

[100] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841* (2019).

[101] Bradley Rey, Bongshin Lee, Eun Kyoung Choe, and Pourang Irani. 2023. Investigating In-Situ Personal Health Data Queries on Smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2023).

[102] Toran Bruce Richards. 2023. Auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous.

[103] Pablo Robles-Granda, Suwen Lin, Xian Wu, Gonzalo J Martinez, Stephen M Mattingly, Edward Moskal, Aaron Striegel, Nitesh V Chawla, Sidney D'Mello, Julie Gregg, et al. 2021. Jointly predicting job performance, personality, cognitive ability, affect, and well-being. *IEEE Computational Intelligence Magazine* (2021).

[104] Kat Roemmich and Nazanin Andalibi. 2021. Data subjects' conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.

[105] Alex Rosenblat, Tamara Kneese, and Danah Boyd. 2014. Workplace surveillance. *Open Society Foundations' Future of Work Commissioned Research Papers* (2014).

[106] Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2013. The usable privacy policy project. In *Technical report, Technical Report, CMU-ISR-13-119*. Carnegie Mellon University.

[107] Norman Sadeh, Jason Hong, Lorrie Cranor, Ian Fette, Patrick Kelley, Madhu Prabaker, and Jinghai Rao. 2009. Understanding and capturing people's privacy policies in a mobile social networking application. *Personal and ubiquitous computing* 13 (2009), 401–412.

[108] Koustuv Saha et al. 2019. Social Media as a Passive Sensor in Longitudinal Studies of Human Behavior and Wellbeing. In *CHI Ext. Abstracts*. ACM.

[109] Koustuv Saha, Ted Grover, Stephen M Mattingly, Vedant Das Swain, Pranshu Gupta, Gonzalo J Martinez, Pablo Robles-Granda, Gloria Mark, Aaron Striegel, and Munmun De Choudhury. 2021. Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–32.

[110] Koustuv Saha, Pranshu Gupta, Gloria Mark, Emre Kiciman, and Munmun De Choudhury. 2024. Observer Effect in Social Media Use. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

[111] Koustuv Saha and Shamsi T Iqbal. 2023. Focus Time for Wellbeing and Work Engagement of Information Workers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems Extended Abstracts*.

[112] Koustuv Saha, Manikanta D Reddy, Vedant das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Raghu Mulukutla, et al. 2019. Imputing missing social media data stream in multisensor studies of human behavior. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 178–184.

[113] Koustuv Saha, Manikanta D Reddy, Stephen M Mattingly, Edward Moskal, Anusha Sirigiri, and Munmun De Choudhury. 2019. LibRA : On LinkedIn based Role Ambiguity and Its Relationship with Wellbeing and Job Performance. *Proc. ACM Hum.-Comput. Interact.* CSCW (2019).

[114] Koustuv Saha, Asra Yousuf, Louis Hickman, Pranshu Gupta, Louis Tay, and Munmun De Choudhury. 2021. A Social Media Study on Demographic Differences in Perceived Job Satisfaction. *Proceedings of the ACM on Human-Computer Ineraction (CSCW)* (2021).

[115] Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. 2018. Employing consumer wearables to detect office workers' cognitive load for interruption management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–20.

[116] N Sadat Shami, Jiang Yang, Laura Panc, Casey Dugan, Tristan Ratchford, Jamie C Rasmussen, Yannick M Assogba, Tal Steier, Todd Soule, Stela Lupushor, et al. 2014. Understanding employee social media chatter with enterprise social pulse. In *Proc. CSCW*.

[117] Mike Sharples. 2022. Automated essay writing: An AIED opinion. *International journal of artificial intelligence in education* 32, 4 (2022), 1119–1126.

[118] Richard Snow. 2013. *I invented the modern age: The rise of Henry Ford*. Simon and Schuster.

[119] Daniel J Solove. 2012. Introduction: Privacy self-management and the consent dilemma. *Harv. L. Rev.* 126 (2012), 1880.

[120] Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science* 48, 2 (2018), 204–231.

[121] Jina Suh, Javier Hernandez, Koustuv Saha, Kathy Dixon, Mehrab Bin Morshed, Esther Howe, Anna Kawakami, and Mary Czerwinski. 2023. Towards successful deployment of wellbeing sensing technologies: Identifying misalignments across contextual boundaries. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–8.

[122] Jim Thatcher, David O'Sullivan, and Dillon Mahmoudi. 2016. Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space* 34, 6 (2016), 990–1006.

[123] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.

[124] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[125] Thea F Van de Mortel et al. 2008. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The* 25, 4 (2008), 40.

[126] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[127] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok K Goel. 2021. Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proc. CHI*.

[128] Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A Scherer, et al. 2016. CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[129] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* (2023).

[130] Angie Zhang, Alexander Boltz, Chun Wei Wang, and Min Kyung Lee. 2022. Algorithmic management reimagined for workers and by workers: Centering worker well-being in gig work. In *CHI Conference on Human Factors in Computing Systems*.

[131] Han Zhang, Vedant Das Swain, Leijie Wang, Nan Gao, Yilun Sheng, Xuhai Xu, Flora D Salim, Koustuv Saha, Anind K Dey, and Jennifer Mankoff. 2024. Illuminating the Unseen: A Framework for Designing and Mitigating Context-induced Harms in Behavioral Sensing. *arXiv preprint arXiv:2404.14665* (2024).

[132] Ou Zheng, Mohamed Abdel-Aty, Dongdong Wang, Zijin Wang, and Shengxuan Ding. 2023. ChatGPT is on the horizon: Could a large language model be all we need for Intelligent Transportation? *arXiv preprint arXiv:2303.05382* (2023).

[133] Kathryn Zickuhr. 2021. Workplace Surveillance Is Becoming the New Normal for US Workers. *Washington Center for Equitable Growth. https://equitablegrowth.org/research-paper/workplace-surveillance-is-becomingthe-new-normal-for-us-workers/. Institute for Research on Labor and Employment University of California, Berkeley* 2521 (2021), 94720–5555.

[134] Manuela Züger, Sebastian C Müller, André N Meyer, and Thomas Fritz. 2018. Sensing interruptibility in the office: A field study on the use of biometric and computer interaction sensors. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.