AI & SOCIETY
https://doi.org/10.1007/s00146-023-01692-3

OPEN FORUM

The unwitting labourer: extracting humanness in AI training

Fabio Morreale[1] · **Elham Bahmanteymouri**[2] · **Brent Burmester**[3] · **Andrew Chen**[4] · **Michelle Thorp**[5]

Received: 31 January 2023 / Accepted: 10 May 2023

Abstract
Many modern digital products use Machine Learning (ML) to emulate human abilities, knowledge, and intellect. In order to achieve this goal, ML systems need the greatest possible quantity of training data to allow the Artificial Intelligence (AI) model to develop an understanding of "what it means to be human". We propose that the processes by which companies collect this data are problematic, because they entail extractive practices that resemble labour exploitation. The article presents four case studies in which unwitting individuals contribute their humanness to develop AI training sets. By employing a post-Marxian framework, we then analyse the characteristic of these individuals and describe the elements of the capture-machine. Then, by describing and characterising the types of applications that are problematic, we set a foundation for defining and justifying interventions to address this form of labour exploitation.

Keywords Immaterial labour · Digital labour · Labour exploitation · Machine learning

1 Introduction

Critical AI scholars have abundantly discussed how AI systems rely on humans. The terms "artificial artificial intelligence" (Stephens 2022) or "human-in-the-loop" (Zanzotto 2019) well summarise this characteristic, which specifically applies to ML, the most dominant AI approach. These systems continuously *learn* from enormous datasets that contain data entries, generally labelled with a "correct answer" that the system can optimise towards. The creation of these datasets often involves micro-tasks that are primarily managed via crowd work (Altenried 2020; Jones 2021; Tubaro et al. 2020; Tubaro & Casilli 2019). As described by Pasquinelli and Joler (2020), "raw data do not exist, as it is dependent on human labour, personal data, and social behaviours that accrue over long period". Scholarly attention has been dedicated to analysing these tasks when performed by undervalued waged workers. However, we find a surprising lack of discussions around those micro-tasks when performed by what we term *unwitting labourers*: individuals who are unaware that typical daily activities they perform online are exploited to train AI datasets. These activities include: adding songs to playlists, accepting/rejecting autocorrect suggestions, offering feedback to a spam filter, filling in a CAPTCHA, uploading photos on digital platforms, and more.

Building upon post-Marxian concepts, we argue that these extractive practices have to be considered forms of labour exploitation as technology companies unilaterally extract surplus value from individuals to train AI. While scholars have previously identified instances in which the product of the labour is behaviour, data, or knowledge (Berardi 2009; Moulier-Boutang 2011; Pasquinelli and Joler 2020; Zuboff 2019), we propose that the exploitative practices of AI training are a distinct category of labour. As we explain in the article, we observe an ongoing systematic *extraction of humanness*. With this concept, we refer to a sociotechnical infrastructure that is systematically deployed to (1) force individuals to generate information on *what it*

✉ Fabio Morreale
f.morreale@auckland.ac.nz

1 Faculty of Creative Arts and Industries (School of Music), The University of Auckland, Auckland, New Zealand

2 Faculty of Creative Arts and Industries (Architecture and Planning), The University of Auckland, Auckland, New Zealand

3 Faculty of Business and Economics (Management and International Business), The University of Auckland, Auckland, New Zealand

4 Faculty of Arts (Koi Tū - the Centre for Informed Futures), The University of Auckland, Auckland, New Zealand

5 Faculty of Arts (Anthropology), The University of Auckland, Auckland, New Zealand

Published online: 17 May 2023

means to be human and (2) extract this information to train AI systems that can produce human-like content. This capturing-machine is purposely developed by AI companies that need human-generated information about an individual's cognitive processes and perceptual performances.

The fact of involuntary extraction of labor value is problematic in itself, as it suggests a form of exploitation paralleling forced labour. However, we do not frame our analysis of the unwitting laborer in terms of labor rights abuse. To do so invites analysis within a liberal paradigm, which places emphasis on the malfeasance of individual 'employers', rather than on the structural features of the digital capitalist economy. Further, 'unwittingness' is a crucial element in our theorisation. The inability of the worker to perceive their work, and, when it is made apparent to them, to elicit recognition of it by institutions wielding power over the discourse of contemporary capitalism, demands an explanation that only critical theory can provide. That is, we must rely on a methodology intended to end self-serving domination over social ontology.

The methodology employed in this research is Political Discourse Theory (PDT) (Glynos and Howarth 2007), an approach to discourse analysis aimed at ontologically problematising, interpreting, and investigating a phenomenon—in this case, individual's work exploitation in AI training. In alignment with Relational Ontology (Rosenberger and Verbeek 2015), and Actor-Network Theory (Latour 2007), PDT explores the essential relations between different actors that play a role in how the phenomenon unfolds (Bahmanteymouri 2021). PDT ontological investigation involves a critique of the existing political, economic, and social relations that create the conditions for a phenomenon to take place and offers a tool to reveal overlooked or hidden relations within a context.

We begin this article by offering some background on the theoretical framing used in our analysis, and in particular, Marxian and Post-Marxian theories of labour exploitation and digital labour exploitation. We then present four case studies representative of the unwitting labourers exploited to train AI systems: reCAPTCHA, content recommendation, content creation, and spam filtering. We then move to identify the characteristic of these labourers against the existing definition of labour and worker. Finally, we discuss the implication of this work and illustrate the concept of humanness extraction in more detail.

2 Methodology

PDT is an ontological approach to discourse analysis used to analyse the nature of social relations, the structures that have shaped these relations, and the nature of economic and social interactions. The approach was initially introduced by Laclau and Mouffe (1985) using Gramsci and Althusser's work to tackle problems of class reductionism and economic determinism. The approach focuses on the questions of those social, cultural, and ideological features that have structured the human subject (Glynos and Howarth 2007). PDT is a post-positivism approach (Glynos et al. 2009), while it suggests using paradigms of the positivistic approach and universal causal laws such as Marxian concepts and theories about capital accumulation and capitalist exploitation, it also emphasises an in-depth analysis of contextualised impacts of meanings, beliefs, and norms as well as ideologies. PDT employs a *retroductive* mode of reasoning. Retroductive inferences primarily focus on inferring what is not observed (Bahmanteymouri 2016: 15) and offer a tool to theorise, explain, and analyse concepts that may seem farfetched or unlikely. It provides an appropriate tool for overlooked or hidden phenomena in theory and empirical fields of social and political sciences. Retroductive reasoning is an effective analytical method for critiquing structures, norms, circumstances, or other actual and real data that have been obscured.

3 Theoretical framework

A key feature of a post-Marxian conception of the economy is that any capitalist economy necessarily produces surplus value (Žižek 2008). Marx explains that surplus value is an excess produced in addition to the original value in the circulation of money–commodity–money (1887, 104). The capitalist produces and owns the surplus value, which allows for the accumulation of capital that is necessary to maintain the capitalist system. Surplus value is enabled by the constant circulation of capital, which can be enabled—among other aspects—by new technologies, new consumeristic ventures, and new things to commodify, as well as new forms of labour exploitation and new social and political relations to help overcome economic downturns (Fuchs 2014; Žižek 2008).

3.1 Marxian and post-Marxian labour exploitation

The necessity logic of constant surplus value creates the conditions for labour exploitation in all forms of capitalism, from industrial capitalism to digital or communicative capitalism and cognitive capitalism (Dean et al. 2006). The most obvious and visible form of exploitation is connected to reducing labour costs, which is needed to maximise the surplus value. Another form of exploitation is alienation, or the ideological dimension of capitalism. Workers become alienated when they lose the object of production and their human value and become, instead, quantitative units of labour without agency. Following Marx, Gramsci (2001)

clarifies how the ideology of the bourgeoisie has shaped the structure and culture in a way that exploitation and alienation are normalised and, in large part, accepted. Marx (1932) describes *alienation* as dehumanised and objectified: rather than selling their product, workers sell their labour and time to create a product. The source of value for the capitalist thus shifts from the human subject—the worker—to the product.

The idea that we will elaborate in this article is that humans unconsciously assume the status of *labourers* when they unwittingly contribute to creating AI datasets. This idea is related to the concept of *immaterial labour* introduced by the Italian Autonomist Marxists, in particular by Lazzarato (1996), who suggested applying the term labour to some activities that are not typically recognised as *work*. Lazzarato's examples are "those activities involved in defining and fixing cultural and artistic standards, fashions, tastes, consumer norms, and, more strategically, public opinion" (ibid.). Similarly, Terranova (2000) defined these as forms of cultural labour which have been "voluntarily channelled and controversially structured within capitalist business practices". As opposed to previous stages of capitalism, in current capitalism, value and surplus value come from "the mind, language, and creativity" (Berardi 2009: 21).

This mode of labour shifts the locus of value creation from manual inputs to behavioural, attitudinal, and imaginative inputs. In the theorising of Autonomist Marxism, immaterial labour enhances the autonomy of workers or at least those whose work is inseparable from digital machines, but this optimism now seems misplaced. Increasingly, the value created by immaterial labour is no less susceptible to appropriation by machine owners than the material labour of classical factory workers. Indeed, as this paragraph was typed, an increasingly shrewd AI built into Microsoft Word offered plausible suggestions for sentence continuation, using the work of millions of writers before us to become more adept at authorship.

3.2 Digital labour exploitation

The original concept of labour was further extended and challenged with the rise of new digital technologies, but the Marxian logic of the worker creating products and subsequently being alienated from the power enabled by those products remains the same. The human-produced information via digital platforms and online services feeds into the power of platform owners (Dean et al. 2006), as the technical infrastructure that collects and funnels data is not owned by those producing the data (Dean 2015).

Gandini (2021) offers an accurate analysis of the genealogy and evolution of the expression "digital labour". While he argues that the expression has now become an empty signifier, we still adopt this expression as a wide-spectrum term that encompasses a variety of relationships between the worker and digital companies. Some forms of digital labour employ a typical employer–employee relationship: so-called platform-labour, i.e. forms of work that are mediated by digital technology (Van Doorn 2017). This form of work has been thoroughly illustrated by Jones (2021), who poignantly wrote that "the poor and dispossessed now unwittingly train the very machines built to … replace their role in the labour process."

Another form of digital labour includes unpaid leisure activities performed on digital platforms (Ritzer and Jurgernson 2010; Postigo 2016). One type of this relationship is user labour on digital platforms and social media, which has traits of Fuchs' (2012) theory of value based on labour time and the Autonomist Marxist's views described above. These unwaged activities, like posting content on social media or providing ratings or signals of good and bad content (e.g. by upvoting or downvoting or liking or favouriting), are considered forms of labour by scholars, including Scholz (2012) and Fuchs (2011): social media users are exploited as they contribute their labour time without reimbursement.

Even before the advent of social media, Terranova (2000) identified a similar form of unwaged exploitation, which she called *free labour*, through voluntary content moderation. At the turn of the century, Terranova recognised that IT companies were commodifying the contributions of community moderators. Chat moderators would "work long hours and love it" (and by doing so, contributed towards AOL-America Online generating at least $7 million a month). Other examples included developers modifying software packages in the open-source community, building virtual spaces on (text-based) virtual reality systems, and individuals participating in mailing lists. This category of Internet users was described as "simultaneously voluntarily given and unwaged, enjoyed and exploited".

Another example of "free labour (produced) by the free multitudes of the internet" is offered by Pasquinelli (2009): human intelligence, collectively created by users that click through Internet hyperlinks, is extracted by the 'parasite' Google to construct and update their proprietary Page Rank algorithm. In more recent years, Ekbia and Nardi (2017) introduced the concept of heteromation to describe computer-mediated mechanisms that extract value "through billions of tiny moments of labour" via active engagement and invisible control. Heteromation extracts economic value from low-cost or free labour in digital technologies, which is often hidden labour: individuals participating in these activities are unaware that their work is transformed into commodities (ibid.).

Despite this growing literature arguing that users' contribution to data creation should be considered labour, we identify the need to specifically analyse *training* data. As we will explain later in the article, the creation of training data needs a unique form of human involvement and specific

exploitative dynamics that, we argue, place this form of data in an ontologically unique space that deserves to be analysed and criticised separately.

4 Case studies

In this section, we present four case studies from a range of different digital applications to clarify how *humanness* is generated by unwitting trainers and extracted by digital companies.

4.1 ReCAPTCHA

ReCAPTCHA is an online challenge-response test first implemented to detect bots and stop them from accessing web pages or digital services. The system asks page visitors to "prove your humanity". This system displays some prompts to the Internet user, usually images but sometimes audio, who has the task of identifying objects or patterns within these prompts. ReCAPTCHA was used to gather enormous amounts of data to train AI, and it was acquired by Google in 2009, just 2 years after the data acquisition function was put in place (O'Malley 2018; Justie 2020). ReCAPTCHA was used to obtain the training data required to digitise the Google Books collection (O'Malley 2018). When an Internet user wanted to access a web service protected by reCAPTCHA, two words in difficult-to-read formats were shown in a form intended to be readable to humans and not robots (Avanesi and Teurlings 2022). Given the pervasiveness of the reCAPTCHA tool, which was used from websites for buying concert tickets to logging onto government services, the practice of reading and transcribing these words became readily accepted by users. By 2011, enough training data had been collected to digitise the entire Google Books archive and 13 million New York Times articles (O'Malley 2018). In 2014, Google began using the system to gather training data for other image-based applications, such as self-driving vehicles or map improvements (ibid.). ReCAPTCHA started presenting users with photographs, and users were asked to identify things within the photographs (Fig. 1). For example, a user might be presented with an image of a road and asked to identify which part of the picture has traffic lights in them.

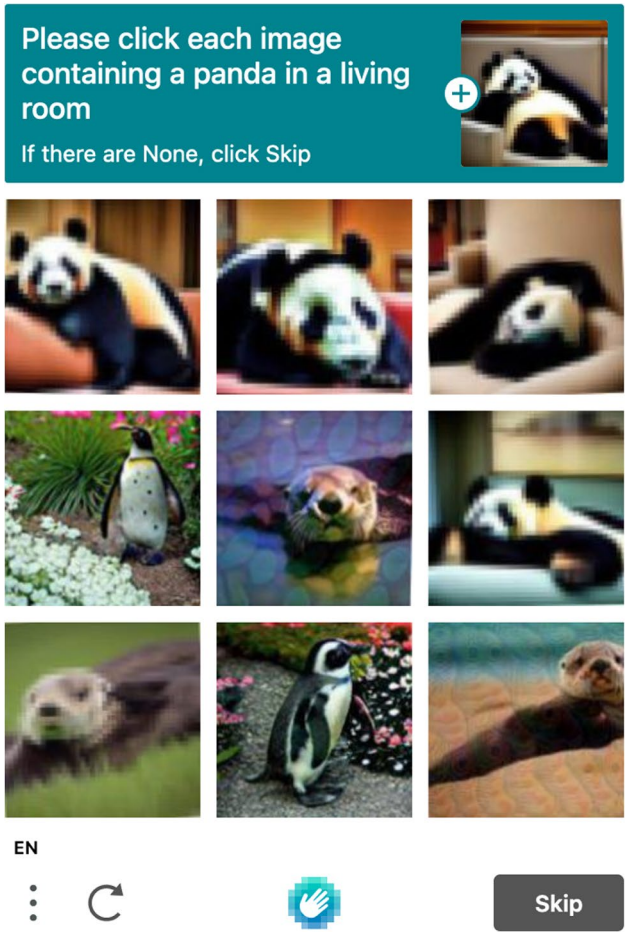

Fig. 1 A typical reCAPTCHA task performed by Internet users every day

4.2 Content recommendation

AI-powered recommendation engines drive many of our daily choices: movies to watch, products to purchase, tourist attractions to see, and travel routes to avoid traffic (or police control). Most of these recommender systems have similarities in that data harvested from unwitting workers is used in their algorithms. In this case study, we explore the case of Spotify, the market-leading music streaming platform, which heavily relies on human-produced data to inform its recommendations. Specifically, it utilises three algorithmic models: collaborative filtering, natural language processing (NLP), and audio analysis (Hodgson 2021; Spotify 2021).

Collaborative filtering is an algorithmic strategy that compares user ratings, interactions, and behaviours against one another to generate new recommendations (Hodgson 2021; Webster et al. 2016). Gustave Söderström, Spotify's Chief Research and Design Officer, explains that when a large group of users put the same songs in the same playlist, this communicates to Spotify that these songs are likely to have something in common (Spotify 2021). This strategy is used to help Spotify create curated playlists for other users. For example, when developing a playlist of *happy songs,* playlist curators can "look for songs that people tend to put on playlists called happy, …[and] we can look at the characteristics of those songs" (ibid.). Söderström acknowledges that Spotify has an "incredibly valuable advantage" of being able to scrape "hundreds of millions of playlists … which

(is) arguably the largest music curation database in history, growing larger every minute to over 4 billion playlists today". He also proudly admits that Spotify has "listening data on how people interacted with the music, lots and lots and lots of listening data" (ibid.).

NLP is utilised by The Echo Nest, a company acquired by Spotify in 2014, to "understand music at scale the way that humans understand music" (ibid.). The Echo Nest employs a set of automated software (bots or crawlers) to determine "how people describe music [and] what are the words people use" (ibid.). These bots and crawlers "go and read the entire Internet and [find] all these blogs, reviews, all sorts of stuff, in order to see how music [is] being described, and then doing natural language processing on top of that" (ibid.). Ajay Kalia, Product Director of Personalisation at Spotify, offers a practical example: if the music of a particular artist continuously gets described as "jangle pop" by online communities, then the algorithm learns to associate "jangle pop" with that artist (ibid.). This method helps Spotify to understand connections between artists and to build an understanding of music in a nuanced and human way.

Finally, audio analysis uses Music Information Retrieval techniques to extract auditory and musical features from songs computationally. These features are compared against other songs to automatically identify similarities and differences between songs.

4.3 Content generation

There has been a lot of hype in recent months around a class of AI systems that generate text, music, visual art, and other forms of output from textual prompts. Increasingly, systems such as GPT-3 (text), Jukebox (music), and Midjourney (images) are producing impressive human-like content.

One example of content generation is Copilot, an AI tool developed by GitHub to autocomplete code as a developer is typing it. Copilot does not autocomplete just one line of code but can provide entire functions (blocks of code) that solve a problem in its entirety. This feature is enticing to developers, because it can save a lot of time by automating low-value programming tasks. The Copilot training set is based on publicly available code and text on the Internet, including public repositories on GitHub. Microsoft, which acquired GitHub in 2018, subsequently integrated Copilot into its widely used integrated developer environment products, such as Microsoft Visual Studio (Wiggers 2022).

Other examples are the art-generating systems like MidJourney, DALL-E, and Stable Diffusion. These AI systems generate images when provided with text prompts, incorporating objects, styles, and context. The training data for these systems comes from datasets of publicly available images on the Internet, such as ImageNet, which are "scraped" (automatically identified and downloaded) from websites like Flickr, YouTube, and social media platforms (Denton et al. 2021). These images include photos taken and art drawn or painted that have been uploaded to the Internet by humans (whether by professionals or everyday people). In many cases, these images are labelled by the image creators themselves, or by other humans through crowdsourcing/crowdworking platforms like Amazon's Mechanical Turk (Mturk 2017).

Similarly, some ML-based solutions to generate new music are currently being developed and have attracted the interest of venture capital investors (Morreale 2021). These systems are trained with a catalogue of existing music, either as raw audio or in notational format, and generate music based on the examples it has "heard". Spotify, which is also active in the AI-music creation space, recently had a job opening for a research scientist working in AI-music creation, which indicated that the new team member would be "using the latest Artificial Intelligence techniques, as well as the huge data sets available at Spotify" (ibid.).

4.4 Spam filtering

Spam filters use a combination of human-defined rules and human-produced training data to identify unsolicited or generally undesirable emails. Some human-defined rules are relatively obvious, such as emails containing the text of known scams or emails coming from email addresses known to send spam. However, spammers are naturally one step ahead of those manually creating these rules. Email providers, thus, must constantly catch up on the latest tactics while ensuring that false positives are minimised (i.e. legitimate emails are not getting caught in spam filters). To address this problem, algorithms have been developed to take a probabilistic approach to spam. They use content and technical meta-data to evaluate the likelihood of an email being spam based on the "features" or "signals" in the email. For example, certain phrases, originating countries, or formatting choices might indicate spam. Individually, these features cannot conclude that an email is spam, but in aggregate, they could lead to high confidence that a certain email is spam.

In order to gather training data, when a human user asks an email client to move an email to spam, the program passes it to the spam filter as a labelled example of spam. Other emails not marked as spam may also be given to the spam filter as examples of "not spam". With millions of users receiving billions of emails daily, a significant corpus of labelled data is available to classify spam filters. From this, these systems can identify patterns and extract features of emails that indicate whether it is spam or not. This form of collaborative filtering is similar to a recommendation engine, although it is used to remove content rather than promote it.

5 Characteristics of unwitting AI trainers

Following a PDT approach, we deploy retroductive inferences to test our theoretical framework—the Marxian theory of surplus value and the post-Marxian immaterial labour—in the cases studies presented above. While different in purposes and potential harm, the selected cases share motivations and processes of gathering data from individuals. This section identifies common features among these different applications and how they interact with unwitting AI trainers. This exercise will ground the discussions that will follow in the final section.

5.1 Unawareness

In all of our case studies, individuals interacting with AI systems are mostly unaware of these data capture practices. When GitHub users started hosting their code on the platform years ago, they were surely not expecting that code then being used to train an AI system. For example, in a demonstration of both how Copilot is built on data generated by humans and the limitations of modern AI systems if a user asks Copilot to autocomplete an "about me" section of a website, it consistently provides content and links to the GitHub and Twitter profiles of a software engineer named David Celis (Gergshron 2021). Celis was not asked permission to use his page content, yet Copilot has decided that his example is canonical. Several lawsuits have since been filed alleging copyright infringement by GitHub and Microsoft in the development of Copilot, despite the code being posted with open-source licenses (Vincent 2022).

ReCAPTCHA users (i.e. virtually all Internet users) are similarly unaware of these extractive practices. The training data gathered by reCAPTCHA has been generated without the users producing it knowing what they are doing or providing informed consent. CAPTCHAs are largely carried out on third-party websites, and, as such, it is not apparent to the user that the micro-task they are performing produces training data that Google can harness in their projects. If the user wants to become aware of how their data and activities are being harnessed, they do not have the option, because information about the algorithmic use of user data is inaccessible (Bartlett et al. 2022). For example, in the reCAPTCHA case, an interested user clicking on the "privacy policy and terms of use" button is redirected to generic Google policies that offer no specific information about the reCAPTCHA itself. As Lung aptly puts it, "reCAPTCHA is opaque in that it is not apparent from the task or context that reCAPTCHA solvers are performing free, menial labour" (2012, 212).

Having characterised labourers solving CAPTCHAs to train AI as *unwitting* does not rule out the possibility that a minority might notice, and resent, the value they are producing with no proper compensation. When this happened in respect of Google's reCAPTCHA service it resulted in an innovative legal challenge. The plaintiffs in *Rojas-Lozano v. Google* (2016) argued they had unknowingly worked for the tech giant by completing the company's reCAPTCHA tests, describing this test as an extraction of free labour. Employment law remains poorly adapted to this scenario, and the case against Google failed. Although the Court agreed that Google had profited, it could not perceive any appreciable loss to users because no reasonable consumer would expect compensation. This outcome aptly exposes the problem we seek to highlight in that the expectations of reasonable consumers enact the prevailing cultural hegemony. Thus, producing consumers, or 'prosumers', are rendered insensible to their own compensable work, *because* it is not compensated. Consequently, despite the notoriety the case earned in the tech community, the general public remains oblivious to the fact that its engagement with CAPTCHA puzzles generates substantial value for Google. (*cf* Cherry, 2016).

As another example from our case studies, music blog and forum users are also unaware that their engagement with these platforms and their reviewing activities are harnessed by The Echo Nest and Spotify to improve their recommendations—particularly if the website they are on carries no mention of Spotify or their branding. While some Spotify users might know that their behaviour is constantly logged and monitored, most are unaware that their interactions with the app—including creating playlists and skipping songs—are all signals used by the company to improve their algorithmic recommendations. This observation resonates with a comment from Morris (2015): "users cannot participate in cloud music without working. " Also, in most cases, artists do not opt-in to having their music used in these AI-music generation training sets. Likewise, users flagging an email as "spam" are not aware that this action will trigger a signal that will be used by the AI to improve the quality of the spam filtering.

Following the theoretical framing employed in this study, these are exemplary cases of workers' labour contributing to creating surplus value for a private company. Whether or not the individual performing these tasks recognises these activities as labour, this type of work is hidden from the labourer (Ekbia and Nardi 2017). These are also the labourers that Marx described: "they do not know it, but they are doing it" (cited in Žižek 2008, p. 16).

The lack of awareness is not an unfortunate accident but rather a systemic feature embedded in these systems to hide the labour. Sadowski (2022) refers to the "structural obfuscation of human labour" when discussing the artful operations performed by AI companies to disguise their real operations. This comment resonates with previous comments from Irani

(2015) when describing the conditions of Mechanical Turk workers. She explains that "hiding the labour" is crucial to the success of data startups to become valued by investors. Similarly, Zuboff (2019) has argued that an essential condition for knowledge production is the ability of surveillance capitalism to evade our awareness.

5.2 Non-consensual labour

Consent is often considered an important part of an agreement to ensure that both parties benefit from a relationship or, if one side does not benefit, that they have at least given permission for the relationship to take place (Eyal 2019, Zwolinski 2018). In traditional working agreements, an employer and employee sign a contract that stipulates the working conditions for the employee and employer, as well as what each can expect of the other. To properly consent, the party is expected to do so voluntarily and understand what they are consenting to (Eyal 2019). In our case studies, the lack of awareness has a direct consequence on consent, as it means that people are not able to give informed consent.

Zwolinksi points out that where consent is not given, the person in question is exploited (2017, 154). In these cases, one party to the agreement has taken undue advantage of the other party.

While in some instances, an individual may give some form of consent (e.g. ticking a box that they have read the Terms and Conditions), they may not be aware of what they are actually consenting to. For example, Google requires the website owner and operator to obtain consent for the data that the reCAPTCHA collects, but they need only do this by asking permission for the website to use Application Programming Interfaces (APIs), which are essentially software interfaces that allow two applications to exchange data (Slack 2022). Thus, although consent is legally obtained, the user is not well informed. To make matters worse, in the reCAPTCHA scenario, the task is presented as a barrier that prevents access to a service that the individual wants to use, essentially forcing the user to complete the task.

Furthermore, employing the Marxian interpretation of exploitation, whether or not a party has consented or not does not change whether or not the relationship can be described as exploitative. Exploitation occurs at any time that there is surplus value produced beyond what society needs. Thus, even if someone has consented to their data being used, they could still be considered exploited.

5.3 Unwaged and uncompensated labour

The exploitation of the unwitting AI trainers in our case studies takes place outside of a wage relationship. Some of these trainers still receive something in exchange for their services. After completing an "obligatory passage-point" (Latour 1994), they gain access to a service, contribute to having a better spam filter, or (supposedly) receive better recommendations.

In other cases, however, these trainers receive nothing in return for their labour. This is the case of music aficionados posting reviews on websites like rateyourmusic.com or commenting on music forums. Their capacity to critically review albums or compare new releases against similar ones is unidirectionally exploited with no recognition by The Echo Nest to improve Spotify's recommendations. Similarly, every individual that uploads art or photos to any publicly available photo-sharing platform (e.g. Google Photos, Flickr, Instagram) is not recognised for their essential work in DALL-E's process of generating new images. While the value of the labour done by individuals may be considered small (or logistically too small to compensate), it is clear that, in aggregate, the private companies leveraging the data can earn significant value and profit.

The concept of an unwaged worker first became apparent in Marxist feminists' analysis of housework and the role of women in a capitalist economy (Federici 2004). Later, Negri (2004) and Smythe (1981) developed the concepts of *social workers* and *audience workers* to argue how the operation of workers has changed because of capitalism. Both images refer to activities such as watching, listening, and reading advertisements, which are necessary for the operation of the capitalist system (Fuchs 2014). As explained by Nardi and Ekbia (2017), "the immense reservoir of labour" that creates content for Instagram "is unpaid but it does not mean that it does not exist". Similar commentaries have been offered by Jones (2021) when discussing digital microwork: "without a wage, one is not quite a worker but a slave, or else surplus", and by Avanesi and Teurlings in their analysis of reCAPTCHA workers (2022): "The absence of a wage relationship does not equal the absence of capitalism". The founder of CAPTCHA, Luis Von Ahn, would surely agree: he has been travelling the world giving seminars called: "How to get people to work for free" (Foley 2014).

5.4 Misappropriation of original intent

Another aspect that brings together all trainers in our case studies is the misappropriation of their original intent. This specific characteristic is one of the most significant differences between the unwitting labourers described in our analysis and those described by previous scholars. In the case of the AOL chat moderators described by Terranova (2000), the specific contribution of the workers was clear and intentional: to moderate chat (thus making it a more enjoyable virtual place to be). In this case, the exploited-intentionality coincided with the exploiter-intentionality. Terranova explained: "Users keep a site alive through their labor, the cumulative hours of accessing the site (thus

generating advertising), writing messages, participating in conversations". Similarly, for gig economy or crowdsourced workers, the purpose of their work is evident. In contrast, the individual's labour is transfigured and repurposed in our case studies.

Wakkary (2021) borrows the concept of *fluid assemblages* introduced by Redström and Wiltse (2018) to explain how technology companies hide layers of functionalities, such as location tracking, to disguise from users "what in fact they are interacting with". The façade of Internet users filling out reCAPTCHA is to prove their humanness, which they unquestioningly perform to have access to a service. By interpreting the images and typing in words, the users think they are proving they were human. In fact, in its original implementation, one of the displayed words was already transcribed and was displayed to test whether the user was a human or machine, but the other was a word that needed to be transcribed and had no correct answer (Justie 2020). Notably, individuals were unaware of which word was the test, so they were forced to complete both to gain access to their desired service. The underlying purpose that they are providing data to a hidden third-party company to train an AI system is structurally hidden.

The intention of an artist when making a song available on Spotify is simple: reach existing and new audiences with their latest release. Spotify rightly fulfils this desire of the musician but at the same time reconfigures the object-song to train their AI in at least two ways. First, using NLP techniques, they acoustically dissect it to find similarities with others and improve Spotify's proprietary recommendation algorithm: "you take a song and then break it down acoustically … chunk it up into little windows, and look at all the characteristics of that song" (Spotify 2021). Second, songs available on music platforms will likely be used as training data to create new music (Morreale 2021).

As another example, GitHub users host their code on the platform for many reasons, including supporting collaboration with other users, cloud storage for their code, and portfolios to demonstrate their capabilities to potential employers and other developers. It is safe to say that no user puts their code on GitHub with a voluntary intent to train an AI model that generates new code. Notably, in the case studies with an element of content generation (such as in AI music generation or Copilot), not only is the purpose of the work transfigured and commodified, but also the system is then used to replace the users who provided the training data in the first place. By adopting this lens, we anticipate seeing novel forms of labour exploitation emerging with an increased dependence on AI.

5.5 The unwitting labourers

Throughout the article, we have described unsuspecting AI trainers as "unwitting labourers". It could be argued that the condition of being unwitting specifically refers to only one of the characteristics identified in this section: unawareness. However, awareness—or wit—is a necessary condition for consent and monetary remuneration: one cannot enter contractual agreements and consent to them without knowing. Linguistic discussions aside, it is only the concurrent existence of all these characteristics that marks the idiosyncratic aspects of the labourer we are describing. Variations of this labour, in which one or more of these characteristics are absent, exist. For instance, some workers are properly contracted and, albeit minimally, remunerated for solving CAPTCHA challenges (Pettis 2022). As another example, some musicians are at least partially knowingly involved in the production of new music AI systems. This is the case of the newly released AI music generation system by Google Research (Agostinelli et al. 2023), whose training included the work of musicians labelling 5521 music examples.

6 Discussion

Moulier–Boutang defined cognitive capitalism as a mode of "capture of gains arising from knowledge and innovation" (2011, p 57). Pasquinelli and Joler (2020) later applied a similar argument to AI, which they describe as an "instrument of *knowledge* extractivism" (italicisation is ours). We argue that, for the category of individuals we have identified in this article as unwitting AI trainers, the object of capture is not simply *knowledge* or *innovation* but rather their *humanness*. With *humanness,* we refer to cognitive processes (e.g. preferences, conceptual associations, identifications, and resolutions) and perceptual performances (e.g. being able to read a word or distinguish a panda from an otter).[1] These processes and performances depict *what it means to be human* and constitute the material with which ML systems are often trained.

In order to enact this process, a technical infrastructure is designed to perform a two-stage process. First, individuals are convinced or forced to perform micro-tasks that (so far) only humans can do. This stage is not always in place as these micro-tasks have already been performed, and the company only has to capture it, such as in the case of The Echo Nest or in the construction of image datasets scraped

[1] This latter category was described by Foley (2014) when discussing Internet users solving reCAPTCHAs. She suggested that psychophysical *perceptual performances* of human bodies have turned into "prosthetics for computation".

from the Internet. Second, these tasks generate information, which is then used to train AI models. This extraction process thus becomes the basis of an artificial know-how, which may be deployed in a myriad of ways by property rights holders to generate rents. Importantly, individuals performing these micro-tasks are unaware of either of these two stages.

Our argument extends the idea of value capture from *immaterial labour* in contracted employment to value capture from individuals that are unwaged, unaware, non-consenting, and whose intention for voluntarily engaging with an activity has been repurposed. In a conventional relationship, both employer and employee understand that at least some proportion of the value produced in labour production is not returned to the labourer. This is the extraction of surplus value that powers capitalism. However, in the scenario we are focused on, the 'employee' is not aware that they are generating value extracted by AI proprietors: they do not appreciate that their immaterial labour has external value and do not use the collective power this affords to make demands of their 'employers'.

The motivations for individuals to engage in these activities are multiple and deserves further scrutiny. Ekbia and Nardi (2017) set out to understand the un-obvious logic for performing volunteer labour in this context. They argue that *predicaments* like separation, precarity, and monotony incite this sort of participation to find relief from everyday struggles. Also, they suggest that this form of participation is encouraged "through an intricate set of mechanisms comprised of social and emotional rewards, monetary compensation, and coercion". When commenting on the exploitative nature of the Weather Channel apps, which "deceptively collect personal data as a business model", Wakkary (2021, p. 197) identified a motivation for engagement in *convenience*. In general, however, humans working for humanness-capturing machines have no option but to interact with these tools to receive certain benefits. This is what Latour (1994) defined an "obligatory passage-point": users have to interact with the tools to be granted access to a specific website or function. An example, as noted by several scholars (Drott 2018; Morreale & Eriksson 2020; Seaver 2022), are Spotify's users: while interacting with the Spotify interface, they constantly feed their preferences and ways of using the app into a proprietary recommendation algorithm. This form of relationship between Spotify and its users is asymmetric, as the user—using the terminology of Marx as cited in Perelman (2000: 30).—is "*compelled to sell themselves voluntarily*". The term *users* is thus not an accurate way of thinking about these phenomena, as they are (also) *unwitting labourers*.

We conclude with a comment on the definition of the individuals participating in AI training. We argue that the linguistic strategy of labelling them as *users* or *customers* is a way of concealing the labour exploitation that is underway. The literature on AI, Human–Computer Interaction, and digital technologies more generally has normalised the term *user*, which has masked a loss of autonomy that is entirely consistent with that experienced by those doing labour. The autonomy these labourers once possessed by virtue of the skills and knowledge they command is taken from them as technology companies appropriate those skills and embody them in their technological products. Once this alienation is complete, artisans become labour units assigned to work with tools made available to them by their employers. Now that this type of labour exploitation has been described and characterised, it can be put into terms that justify and describe the change. Our societies have allowed a business model founded on exploitation to proliferate, and it should be in the interests of governments to reorient the balance of power. Our future work includes further developing the theory underpinning our arguments, more concretely demonstrating the harm from this exploitation, and making the case to policymakers and the public that appropriate interventions are necessary to mitigate this unbridled accumulation of value and power.

Acknowledgements We thank our many colleagues who have participated in conversations and workshops that helped us shape our thinking and clarify our arguments. In particular, we thank Briony Blackmore and Matt Bartlett for their significant contribution to our work.

Author contributions The project originated from discussions between FM and BB, but all authors contributed to the project's conception. The theoretical framework was mostly provided by EB, BB, and FM. The case studies were mostly investigated by AC (reCAPTCHA, CoPilot, spam filtering) and FM (Spotify, Midjourney). The methodological aspects connected to PDT were led by EB. The write-up of the manuscript was led by FM and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. The University of Auckland Transdisciplinary Ideation Fund provided research funding for this project (Grant 3725387).

Data availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organisation or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

Ethical approval Ethics approval is not required for this project.

References

Agostinelli A, Denk TI, Borsos Z, Engel J, Verzetti M, Caillon A, Huang Q, Jansen A, Roberts A, Tagliasacchi M, Sharifi M (2023) Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325

Altenried M (2020) The platform as factory: crowdwork and the hidden labour behind artificial intelligence. Cap Class 44(2):145–158

Avanesi V, Teurlings J (2022) "I'm Not a Robot," or am I?: Micro-Labor and the Immanent Subsumption of the Social in the Human Computation of ReCAPTCHAs. Int J Commun 16:1441–1459

Bahmanteymouri E (2016) An ontological investigation of urban growth management policies under neoliberalism. PhD Thesis, The University of Auckland. ResearchSpace@Auckland

Bahmanteymouri E (2021) A Lacanian understanding of urban development plans under the neoliberal discourse. Plan Theory 20(3):231–254

Bartlett M, Morreale F, Prabhakar, G (2022) Analysing Privacy Policies and Terms of Use to understand algorithmic recommendations: the case studies of Tinder and Spotify. J Roy Soc New Zealand, 1–14

Berardi F (2009) The soul at work: from alienation to autonomy. MIT Press, Cambridge

Cherry MA (2015) Beyond misclassification: the digital transformation of work (February 18, 2016). Compar Labor Law Policy J 37:577

Dean J, Anderson JW, Lovink G (eds) (2006) Reformatting politics: information technology and global civil society. Routledge

Dean J (2015) Technology: the promises of communicative capitalism. In: Reclaiming Democracy (pp 58–84). Routledge

Denton E, Hanna A, Amironesei R, Smart A, Nicole H (2021) On the genealogy of machine learning datasets: a critical history of ImageNet. Big Data Soc 8(2):20539517211035956

Drott EA (2018) Music as a technology of surveillance. J Soc Am Music 12(3):233–267

Ekbia HR, Nardi BA (2017) Heteromation, and other stories of computing and capitalism. MIT Press

Eyal N (2019) Informed consent. Stanford Encylopedia of Philosophy. https://plato.stanford.edu/entries/informed-consent/. Accessed 30 Jun 2022.

Federici S (2004) Caliban and the Witch. Autonomedia.

Foley M (2014) "Prove You're Human": fetishizing material embodiment and immaterial labor in information networks. Crit Stud Media Commun 31(5):365–379

Fuchs C (2011) An alternative view of privacy on facebook. Information 2:140–165

Fuchs C (2012) The political economy of privacy on Facebook. Television & New Media 13(2):139–159

Fuchs C (2014) Digital labour and karl marx. Routledge

Gandini A (2021) Digital labour: an empty signifier? Media Cult Soc 43(2):369–380

Gershgorn D (2021) GitHub's automatic coding tool rests on untested legal ground. The Verge. https://www.theverge.com/2021/7/7/22561180/GitHub-copilot-legal-copyright-fair-use-public-code

Glynos J, Howarth D (2007) Logics of critical explanation in social and political theory. Routledge

Glynos J, Howarth D, Norval A, Speed E (2009) Discourse analysis: varieties and methods

Gramsci A (2001) Selections from the prison notebooks of Antonio Gramsci. Electronic Book Company

Hodgson T (2021) Spotify and the democratisation of music. Pop Music 40(1):1–17

Irani L (2015) Difference and dependence among digital workers: the case of Amazon Mechanical Turk. South Atlantic Quart 114(1):225–234

Jones P (2021) Work without the worker: labour in the age of platform capitalism. Verso Books

Justie B (2020) Little history of CAPTCHA. Internet Histories 5(1):30–47. https://doi.org/10.1080/24701475.2020.1831197

Laclau E, Mouffe C (1985) Hegemony and socialist strategy. Verso, London

Latour B (1994) On technical mediation—philosophy, sociology, genealogy. Common Knowl 3(2):29–64

Latour B (2007) Reassembling the social: an introduction to actor-network-theory. Oup Oxford

Lazzarato M (1996) Immaterial labor. Radical Thought Italy 1996:133–147

Lung J (2012) Ethical and legal considerations of reCAPTCHA. In: 2012 Tenth Annual International Conference on Privacy, Security and Trust

Marx K (1887) Capital: a critique of political economy, book one. In: Blunden A, Clayton C (2008), Harris M (2010) (Eds) The process of production of capital. Retrieved from http://www.marxists.org/archive/marx/works/download/pdf/Capital, 1

Marx K (1932) Economic and philosophic manuscripts of 1844. https://marxists.org/archive/marx/works/download/pdf/Economic-Philosophic-Manuscripts-1844.pdf on 24 Jan 2023.

Morreale F (2021) Where does the buck stop? Ethical and political issues with AI in music creation. Trans Int Soc Music Inform Retriev 4(1):105–113

Morreale F, Eriksson M (2020) "My Library Has Just Been Obliterated": producing new norms of use via software update. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp 1–13)

Morris JW (2015) Selling digital music, formatting culture. In: Selling digital music, formatting culture. University of California Press

Moulier-Boutang Y (2011) Cognitive capitalism. Polity

MTurk (2017) Tutorial: How to label thousands of images using the crowd. https://blog.mturk.com/tutorial-how-to-label-thousands-of-images-using-the-crowd-bea164ccbefc - accessed on 24 Jan 2023

Negri A (2004) Time for revolution. A&C Black

O'Malley J (2018) Captcha if you can: how you've been training AI for years without realising it. Tech Radar. January 2018

Pasquinelli M (2009) Google's PageRank algorithm: a diagram of cognitive capitalism and the rentier of the common intellect. Deep Search 152–162

Pasquinelli M, Joler V (2020) The Nooscope manifested: AI as instrument of knowledge extractivism. AI Soc 1–18

Perelman M (2000) The invention of capitalism: classical political economy and the secret history of primitive accumulation. Duke University Press

Pettis BT (2022) reCAPTCHA challenges and the production of the ideal web user. Convergence. https://doi.org/10.1177/13548565221145449
Postigo H (2016) The sociotechnical architecture of digital labor: converting play into YouTube money. New Media Soc 18(2):332–349
Redström J, Wiltse H (2018) Changing things: the future of objects in a digital world. Bloomsbury Publishing
Ritzer G, Jurgenson N (2010) Production, consumption, prosumption: the nature of capitalism in the age of the digital "prosumer." J Consum Cult 10(1):13–36
Rojas-Lozano v. Google Inc (2016) 159 F. Supp. 3d 1101
Rosenberger R, Verbeek PP (2015) A field guide to postphenomenology. In: Rosenberger R, Verbeek P-P (eds) Postphenomenological Investigations: essays on HumanTechnology Relations. Lexington Books, Lanham, pp 9–41
Sadowski J (2022) Planetary Potemkin AI: The Humans Hidden inside Mechanical Minds. Digital Work in the Planetary Market, 229
Scholz T (ed) (2012) Digital labor: the Internet as playground and factory. Routledge
Seaver N (2022) Computing taste: algorithms and the makers of music recommendation. University of Chicago Press
Slack C (2022). "Privacy Policy for ReCAPTCHA. Free Privacy Policy. https://www.freeprivacypolicy.com/blog/recaptcha-privacy-policy/. Accessed on 30/08/2022
Smythe DW (1981) Dependency road. Ablex, Norwood
Spotify (2021) Human vs machine—spotify: a product story. https://open.spotify.com/episode/0T3nb0PcpvqA4o1BbbQWpp
Stephens E (2022). The mechanical Turk: a short history of 'artificial artificial intelligence. Cultural Studies, 1–23
Terranova T (2000) Free labor: producing culture for the digital economy. Social Text 18(2):33–58
Tubaro P, Casilli AA (2019) Micro-work, artificial intelligence and the automotive industry. J Indust Bus Econ 46(3):333–345
Tubaro P, Casilli AA, Coville M (2020) The trainer, the verifier, the imitator: three ways in which human platform workers support artificial intelligence. Big Data Soc. https://doi.org/10.1177/2053951720919776
Van Doorn N (2017) Platform labor: on the gendered and racialised exploitation of low-income service work in the "on-demand" economy. Inf Commun Soc 20(6):898–914
Vincent J (2022) The lawsuit that could rewrite the rules of AI copyright. The Verge. https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data. Accessed on 29 Jan 2023.
Wakkary R (2021) Things we could design: For more than human-centered worlds. MIT press
Webster J, Gibbins N, Halford S, Hracs B J (2016) Towards a theoretical approach for analysing music recommender systems as socio-technical cultural intermediaries. In: Proceedings of the 8th ACM Conference on Web Science pp 137–145
Wiggers K (2022) Copilot, GitHub's AI-powered programming assistant, is now generally available. TechCrunch. https://techcrunch.com/2022/06/21/copilot-githubs-ai-powered-programming-assistant-is-now-generally-available. Accessed on 29 Jan 2023.
Zanzotto FM (2019) Human-in-the-loop artificial intelligence. J Artif Intell Res 64:243–252
Žižek S (2008) The sublime object of ideology. Verso, London
Zuboff S (2019) The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019. Profile books
Zwolinski ZM (2018) Exploitation and consent. In: Müller A, Schaber M (eds) The Routledge handbook of the ethics of consent. Routledge, New York, pp 153–163