THE UNIVERSITY OF CHICAGO

Amazon Mechanical Turk and Machine Learning Algorithms:

A Computational Content Analysis of Digital Crowdlabor in the Production of Artificial Intelligence

By

Alexander C. Melamed

February 2021

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts Program in the Social Sciences

Faculty Advisor: Amit Anshumali

Preceptor: Sophie Fajardo

Abstract

Among the many uses for Amazon's *Mechanical Turk* digital crowdlabor platform, the production of artificial intelligence (*AI*) – specifically, machine learning algorithms – is one of the most popular, and certainly among the most idiosyncratic. Much has been theorized about the relationship between AI and human labor from various sociological perspectives, but no studies have directly addressed the question of why digital crowdlabor appears so suited to the production of machine learning algorithms. This article seeks to investigate that question by using computational content analysis techniques to develop a hermeneutic framework through which the process of AI production via Mechanical Turk can be understood, given the available data from the digital crowdlabor platform.

The data analyzed by this study is taken from a corpus of digital text composed of the available descriptions and metadata from 71008 unique *human intelligence tasks* (*HITs*) published on the Mechanical Turk (*mTurk*) website. The mTurk HIT listings comprise a virtual marketplace where private employers (*Requesters*) place open bids for the collective cognitive labor of individual mTurk workers (formally styled as *Workers* and known informally as *Turkers*) which is computationally agglomerated by the platform into a dataset for the Requester. In addition to the programming of the algorithm itself, the process of producing a machine learning (*ML*) algorithm necessarily requires a corpus of relevant data which has been digitally formatted and annotated such that the algorithm can be refined using it. The production of these corpora is only feasibly accomplished by the collaboration and coordination of many individuals – precisely what digital crowdlabor platforms (*DCPs*) like mTurk are designed to provide.

It is concluded from the results of this study that what makes DCPs not merely suitable but also highly profitable for producing ML is, at least in part, that they semiotically equate the surplus-value of the labor-power of human workers with the use-value of the computational machinery they own and operate. In other words, this semiotic equation of cognitive labor with computational value allows DCPs such as mTurk to extract profits from not only human workers but also from the digital technology those workers own and use in their work, all while eliding that very distinction. In this sense, the production of ML algorithms via digital crowdlabor can be theorized as uniquely suitable to the extent that DCPs can legally capitalize upon the productive potential of crowdsourced computational machines, which would otherwise need to be invested by the Requesters themselves. By facilitating this economic alienation of workers from their property, the productive relations of mTurk have positioned the production of ML algorithms as a new and effective means for extracting increased profits, one unique to digital capitalism.

1. Introduction

The 21st century has witnessed the ascendance of digital computers and the internet to a level of global ubiquity deemed sufficient by many scholars to constitute a qualitatively new technological basis for production and commerce, as distinguished from the preceding industrial mode characterized by centralized factories and formal employment. Among the most promising products of the *post-industrial* digital economy are the menagerie of artificial intelligence programs that are revolutionizing contemporary global society -- most promising among them being the family of programs known as machine learning algorithms. These algorithms are often produced by private contracting of digital crowdlabor through DCPs such as mTurk. Not coincidentally, mTurk was first born as a solution to the failure of an Amazon project to develop a machine learning algorithm, one that could automatically scan third-party product listings on the company's online store and flag duplicate items for removal. When the necessary labor required to produce such an algorithm proved unsustainable under traditional software development procedures, alternative methods for sourcing and mediating large amounts of digital labor were investigated. Pitched by Jeff Bezos internally to the company and again to the world during its 2006 public launch as "artificial artificial intelligence" (Aytes 2013, 80) and "humans-as-a-service" (Irani 2015, 225), Amazon Mechanical Turk has always embodied the central contradiction of AI -- unprecedented technological advancement, achieved at the customary cost of human exploitation.

This study aims to answer the question of what makes mTurk – and thereby DCPs in general – so suited to the production of machine learning algorithms. A review of the literature on digital labor and artificial intelligence failed to uncover any prior studies that investigate the specific productive relationship between DCPs and ML algorithms. It is apparent from the history of the development of mTurk that the relationship between crowdlabor and ML thereby established has been motivated by the profit imperative. The commodification of digital crowdlabor in the production of ML is what underpins its profitability, and thereby its success where prior attempts had failed. The elucidation of that commodification in systematized terms would entail significance across the fields of digital labor and artificial intelligence, perhaps even helping to resolve current disputes regarding the nature of AI itself.

This article posits that the HIT work performed by crowdworkers for ML platforms such as mTurk can be categorized as *immaterial* and *cognitive* labor. The term cognitive labor is typically used to describe a subset of immaterial labor involving primarily mental tasks, specifically identified by some scholars with the production of machines, both material and social (Aytes 2013, 94). Digital crowdlabor, as a subset of cognitive labor, becomes comprehensible then in terms of the production of digital machines through the cognitive labor of crowdworkers. Immaterial labor, meanwhile, is more generally typified as that labor which "produces the informational and cultural content of the commodity" which could be anything, including but not limited to a machine of some sort (Lazzarato 2014, 32). The much-heralded post-industrialization of the global economy during the second half of the 20th century has been characterized by a new affirmation of the value of such immaterial labor, ushered in by the ascendency of digital technology as the product and means of labor both. Some scholars of the post-industrial economy have theorized that this ascendence portends a world-historic transformation for global capitalism, or even constitutes the prelude to some entirely new mode of production (Hardt and Negri 2012). For other scholars however, the development of new and increasingly profitable means of extracting surplus-value from immaterial labor -- even in the context of an ostensibly post-industrial economy -- remains fully comprehensible within the extant operational logic of modern capitalism. Of these two positions, it is the latter which is broadly adopted by this article. For all its technological marvels, the shift towards immaterial labor in has not yet coincided with any revolutions successful in overturning the logics of capitalist exploitation. Fueled by increasingly powerful digital technology, it is precisely those logics in fact which have become most elided even as they are heightened by the consolidation of international finance, the rarefication of working-class solidarity, and the acceleration of both those trends.

That said, while HIT work is still understood as essentially congruous with the basic blueprint for commodified labor-power established by the historical trajectory of industrial capitalism, the instrumentalization of digital crowdlabor in the production of ML algorithms is also understood as a meaningfully new tangent branching off from that trajectory. The propensity of quantity to become quality should not be neglected, and while massive reserves of cheap labor -- even digital labor -- have been available to capitalists long before the DCP was ever devised, the coordination and concentration of that labor has been greatly accelerated by the advent of new means of immaterial labor such as mTurk. By algorithmically objectifying the cognitive labor of numerous digital crowdworkers into the form of a commodified artificial intelligence that is alienated even further from human labor by being socially construed as artificial i.e., *inhuman* (Dyer-Witheford, Kjøsen, and Steinhoff 2019). Thus, the question is raised of whether DCPs are so uniquely suited to ML production by virtue of the technology which facilitates this commodification, or the manifestation of some novel form of economic exploitation unique to DCPs but unrelated to said technology.

This study hypothesizes, that in addition to the heightened extraction of value from the labor of workers that it facilitates, the mTurk DCP also extracts an additional source of value from computational fixed capital by requiring digital crowdworkers to provide the tools of their trade at their own expense. Analysis of the data collected from the mTurk DCP supports this hypothesis. A pattern is identified wherein the highest paying HITs require workers to provide and utilize digital peripheral equipment, such as webcams and digital recording devices, whereas lower paying HITs only require the basic functions of a personal computer. In both cases value is extracted from the computational property in use, but the sharp increase in compensation that coincides with an increase in the technologization of labor required corroborates the theory that this aspect of the DCP's productive relations is a significant economic factor.

1.1. Research Significance

It has been previously established that digital crowdlabor and machine learning algorithms are closely interrelated in academia and industry (Quinn et al 2010; Reynolds, Kontostathis, and Edwards 2011; Le et al 2010; Laws, Scheible, and Schütze 2011; Lawson et al 2010; Hara et al 2014; Zhao, Sukthankar, and Sukthankar 2011; Mitra, Hutto, and Gilbert 2015; Sriraman, Bragg, and Kulkarni 2017; Callison-Burch and Dredze 2010). Therefore, the study of these interrelated topics is of interest in many areas of the social sciences, from political-economy to science studies.

While alternatives do now exist, mTurk is the largest and most widely recognized digital crowdlabor platform publicly available and it has remained so ever since its 2006 launch. As the de facto industry standard, mTurk represents an ideal source of data from which to draw insights potentially generalizable across the entire sector of digital crowdlabor. By extension of this rule, the analytical results of this article may provide insights horizontally applicable to those industries and institutions that employ digitally crowdsourced microlabor, in addition to furthering the general academic understanding of the relationship of crowdlabor to artificial intelligence. Regardless, it seems inevitable that the market for digital crowdlabor will continue to expand alongside any projected increases in demand for commodified artificial intelligence.

1.2. Research Implications

In practical sense, a computational content analysis of mTurk HITs could benefit the conditions of mTurk Workers themselves, given the fact that many of them already employ third party programs for automatically filtering and sorting HIT listings based on such information (Irani and Silberman 2013, 611; Gupta et al 2014, 3-4). By contributing to the existing knowledgebase of that information, this research could help improve the ability of mTurk Workers to negotiate in the digital crowdlabor market more optimally. Some may look at the emergence of such techniques in combination with a highly deregulated and virtually decentralized marketplace and consider it self-evident that DCPs empower participating individuals as competitive economic actors. This article will challenge such interpretations by following from the premise that workers engaging in any productive activity do so definitionally as purveyors of their labor-power. Therefore, workers are presumed to be ultimately beholden to the mandates of capitalism, as manifested in the social and material relations structuring the means of production under which they engage in the activity of labor. As one such means, computational platforms for the digital crowdsourcing of cognitive microlabor can only be considered empowering along a logarithmic curve, as David Golumbia describes it, where "the relatively disempowered are certainly enfranchised [but] the powerful are themselves radically reinforced" within their existing positions of political-economic power (Golumbia 2009, 145). Any progress made herein towards demystifying the relations of digital crowdlabor and digital capital would, ideally, contribute to the leveling of that curve.

2. Literature Review

Labor scholars studying contemporary forms of non-traditional or *informal* employment have generally placed digital crowdlabor platforms within that category. This is because the workers involved therein are not legally qualified as employees but rather independent contractors and as such must "use their own equipment, cover operating costs, handle their own accounting and taxes, and provide their own insurance" (Vallas 2019, 49). Despite the common idealization of independent contract work as flexible and discretional, the computational nature of digital crowdlabor platforms allow for certain techniques for intensifying managerial control which appear to be the result of having adapted the neoliberal outsourcing paradigm to an international, digitally networked cognitive labor community. As such, these techniques seemingly straddle the historical divide between industrial and post-industrial labor management.

2.1. Digital Crowdlabor and the Informal Economy

The *fragmentation of tasks* model is one example of the aforementioned managerial techniques unique to digital crowdlabor. As a characteristic of the technological relations of production, the fragmentation of tasks model describes the structural means by which microlabor performed by multiple workers in the completion of a particular HIT is computationally agglomerated and homogenized for the purpose of standardization as a coherent dataset. As a characteristic of the social relations of production, the fragmentation of tasks also serves to systemically alienate those same workers from eachother, greatly inhibiting collective action and other forms of solidarity by the same mechanism that facilitates the concentration of their collective labor's products (Ayes 2013, 79).

Even the extremely low average rates at which Turkers are typically compensated for their labor-time are subject to certain forms of managerial arbitration. Once the HIT work performed by a Turker has been submitted for review, it becomes the legal intellectual property of the Requester, who is under no obligation not to simply reject the work as unsatisfactory and thereby recusing themselves from having to pay for it. Further compounding this issue, HIT work submissions are often flagged for approval or rejection by semiautonomous algorithmic protocols. These algorithmic methods are no more arbitrary even than the most common manual methods, some of which involve screening individual Worker accounts for competence by issuing them pre-solved tasks disguised in the form of normal HITs as tests, only presenting the real contracts to those who pass the tests. Another approach known as "majority rule" involves mass-issuing the same unsolved HIT to numerous workers; the most common responses are designated correct and those who provided them greenlit, while the others are blacklisted or denied payment (Irani 2015). There is a process by which workers whose work has been rejected by the contractor can request an appeal from them through mTurk, but there are no conventions compelling the Requester to reply. These appeals are generally unread by the Requesters, who interpret them as indirect indicators of the performance of their algorithmic methods for determining the correctness of received HIT work responses. That is, if an unusually large number of appeals are being issued in regard to a particular HIT, that is taken as a sign that the algorithmic management of that HIT potentially needs revision. However, even when revisions are made, the appeals themselves are not often resolved (Ibid).

2.2. The Subjectivities of Digital Crowdworkers

Each of the individualizing and alienating technological and social relations of HIT work described in the preceding subsection reflect and facilitate in some way the deterritorializing effects of *labor arbitrage* – i.e., the dissociation of workers from their locally-situated identities and reterritorialization as transnational subjects under the platform company (Ellmer 2015, Aytes 2013). In this context, the willing self-subjectivation of mTurk workers as *Turkers* (an epithet coined by the community) can be understood psychoanalytically. From investigation into "the technical means by which diverse workers are rendered into computational resources" for the profit of digital crowdlabor Requesters, Lily Irani concludes that such novel subjectivities produced by DCPs extend to the Requesters as well as the Workers. The self-imagination of the latter class as "programmers, entrepreneurs, and innovators" is discursively validated and reproduced by the DCP systems which computationalize the labor they can then effortlessly expropriate (Irani 2015, 226-229). This subjectivation is economically sustained also, since sourcing labor through DCPs instead of conventional employment legally obscures the real labor costs involved. The reduced external visibility and increased internal atomization of the workforce involved effectively frees Requesters to experiment with risky or unethical strategies of labor management and market expansion. The invisiblization of labor and the reduction of investment risk has the cumulative benefit for enterprising Requesters of inflating the value of their cutting-edge technological ventures, which "attract more generous investment terms when investors perceive them as technology companies rather than labor companies" (Irani 2015, 230-232).

2.3. The Historical Genealogy of Digital Crowdlabor

Lily Irani draws connections between the invisibilization-computationalization continuum described in the previous subsection and the research of feminist scholars into the similar erasure of the "creativity and improvisation" involved in so-called menial labor, historically effected along gendered lines whereby women are disproportionately targeted (Irani 2015, 231). Viewed through the lens of this historical connotation, mTurk is revealed to be hardly the first example of exploitatively arbitrated crowdsourced microlabor. In 1985, the telemarketing agency DialAmerica contracted American homeworkers (an occupation overwhelmingly gendered as feminine) *en masse* with the task of independently confirming the phone numbers of 500 private individuals at a compensation rate of 5-10 cents USD per number (Felstiner 2011). Eventually DialAmerica was brought to court over this matter by the US Secretary of Labor, but the case was dismissed on the ruling that the homeworkers legally qualified as independent contractors who "'had the freedom to work at any time and for as many hours as they desired'" as they were not working under direct management by the company (Felstiner 2011, 127).

In many ways, mTurk is to the internet today what the DialAmerica homework campaign was to the telephone in 1985. The main distinction between the Dial America and Amazon Mechanical Turk business strategies relates to the qualitative difference between the telephone and the digital computer which has enabled the mTurk model to benefit from much greater economies of scale than that of the DialAmerica strategy. Fortunately, Steven P. Vallas in "Platform Capitalism: What's at Stake for Workers?" makes a strong case for the "retail revolution" of the 1980s qualifying as another point on the trajectory of exploitative "platform economy" business strategies, historically located between the DialAmerica and mTurk models (Vallas 2019, 49-50). Vallas describes the retail revolution as a phenomenon enabled by the neoliberal economics of the Reagan-Thatcher era whereby "upstart[1] discount firms such as Walmart and Target […] undermined the power of domestic manufacturers and incumbent department stores" by leveraging cutting-edge logistic through methods directly inspired by military

[1] The term "upstart" as used here might seem to imply that Walmart and Target were newly incorporated companies at this time, when they were in fact founded in 1962 and 1902, respectively. This contradiction between the historical pedigree of corporate capital and its apparent nimbleness in 'revolutionizing' its own industry when strategically adopting advanced technologies illustrates the power of technological development as a qualitative factor in capitalist production, suggesting also that the investment cost required can be a barrier so steep as to require the accumulation of said pedigree before being cleared, else it would have been done sooner.

techniques (Vallas 2019, 52-53). At every level, the economic techniques of the retail revolution relied on computational technologies ranging from cash register bar-code scanners to satellite communication networks, which Vallas credits with finally dispelling the notion that information (logistical information specifically in this case) could not be the deciding factor in a firm's profitability, as opposed to tangible capital assets (Vallas 2019).

However, it is difficult to contest that every one of these computational technologies does still represent some form of capital in its own right. The valorization of isolated abstract information captured and transmitted by a satellite ignores the materiality of not just that information, but of value itself as a product of the physical contributions of human labor-power, whether tangibly objectified in the form of the satellite or virtually in the form of digital information. In their introduction to *The Technological Economy*, editors Slater and Barry dispute the conventional dichotomy made so often within the social sciences between "social and material objects" as easily problematized when the origination of those objects as scientific-technological products is considered (Slater and Barry 2005, 6). This reframing thus construes the information object as "informed" i.e., one produced both by and within a particular "informational environment" which is responsible in a productive sense for the very informational character of the object itself (Ibid). Inversely, information in the abstract is only understandable as such, Slater and Barry argue, to the extent that its embeddedness within the complex relations of material objects can be recognized and understood. This dialectical mode of sociological meaning-making is thus regarded as the best means for resolving the apparent contradictions which emerge from investigations of informational objects which would otherwise seek to classify them as wholly either material or social.

If the informational meaning of an object which is understood sociologically as such is necessarily attributable to it having been produced and imbued with that meaning by and within an informational environment, and that informational aspect is not rendered intelligible until properly located within the social-material relations of the environment in which it is analyzed, then the question arises of what if any transformation occurs at the informational level during the migration of the object out of its original environment and into the wider world, where it is subsequently embedded and examined. This question of how to understand the peregrinations of informational objects between production and circulation appears prominently throughout the works of Karl Marx, including a section of the *Grundrisse.* There he asserts that all machines produced under capitalism are by definition informed, in the sense that their very existence is entirely predicated on the preexistence of the specific human knowledge required for their material construction and social signification. Thus, before a satellite or its transmissions can come into existence let alone precipitate surplus-value within a market, there must exist a broad social knowledgebase capable of imagining them – what Marx terms the "general intellect" –

which must then be realized as a productive paradigm under which raw materials are transformed by labor into "*organs of the human brain, created by the human hand;* the power of knowledge, objectified" (Marx 1973, 706).

In recognizing the inherent *informedness* of the knowledge and communications technologies which allowed DialAmerica, Walmart, Target, and Amazon to revolutionize their respective industries, the significance of the general intellect is made all the more apparent. Those pioneers of industry did not invent the technologies in question *ex nihilo*. In fact, the national telephone grid, pre-internet digital logistics, and internetworked computational machines can all be confidently described as being firmly embedded within the social and material infrastructure of the capitalist world at the time of their utilization in each case. In a following section of the *Grundrisse*, Marx adds that major innovations in "*fixed capital"* (i.e., material assets actively invested and consumed in productive processes) indicate the advancement not only of the general intellect but also of the broad accumulation of capital itself within a society, such that there must exist a general degree of productive efficiency sufficient to have yielded a "relative overabundance […] specifically, a level directly related to the transformation of circulating capital [i.e. capital not yet consumed or invested in production] into fixed capital" (Marx 1973, 707). That is to say, advanced technologies employed in industry only yield superlative profits when the profits being extracted from labor-power operating under lower conditions of technological advancement are already great enough so as to compensate for the initial investment cost, and subsequent consumption, of those advanced productive technologies. That each of the case studies examined in this section of the article emerged from a historical context wherein the technologies they capitalized upon were already widely and deeply institutionalized is no coincidence. It is argued therefore, in accordance with Marx, that such a level of technological saturation across the general intellect of a society is preconditional to the profitable utilization of advanced productive technologies under capitalism – including the production of machine learning algorithms using digital crowdlabor platforms. The point here being that the evident success of digital crowdlabor should be taken as indication that the general intellect of global society in the current historical moment must be sufficiently advanced and appropriately formed such that any given individual human finding themselves in the position of a Turker, would be in effect always-already capable of objectivating their cognitive labor into the informed object of the ML algorithm commodity.

By taking advantage of the matured mid-late 20th century American home telephone market to contract homeworkers at maximal scale and minimal cost, DialAmerica were able to effectively commandeer the national telephone grid as their own factory floor but without the regulatory baggage. The retail revolution would not have been possible if the consumer trends and commodities markets that Walmart and Target so effectively capitalized upon had not already been rendered into consistent digitally-readable forms by the national standardization of retail barcodes and the international conventionalization of commercial satellite transponders, for instance. In each case the effective value imbued by the respective informed technologies being leveraged is evidently greater than the sum of their material parts, but that additional value is not imparted by dint of that informational aspect itself. Rather, in such cases, the utility of advanced informational technologies lies in the ease with which they can be reappropriated by private corporations and translated from circulating capital into fixed capital.

What must be considered when reading Marx into the context of contemporary technologies is that, at the time of his writings, the concept of advanced productive technologies was epitomized in the form of the industrial machine – a commodity which could by necessity only be owned let alone put to productive use by a capitalist inside a factory. Today, by contrast, machines are everywhere in the form of digital computers between which information and processing power are perpetually circulating via the internet. The key distinction here is that, unlike industrial era factory machinery, the vast majority of computational commodities today are held as personal rather than private property or in the case of the internet as essentially public infrastructure; in other words, circulating rather than fixed capital. In the formulation related above, Marx likely had in mind forms of circulating capital such as those usually required for social reproduction (e.g., food, clothing, or any other commodity necessary to sustain a worker in their labor) which would be transformed into fixed capital during the process of accompanying the worker in their work and being consumed as such. However, if the productive potential of non-privately owned computational machines as circulating capital is considered under this same economic framework, the revolutionary profitability of digital crowdlabor and its precedents is clearly explicated.

2.4. Theorizing Human Labor in the Age of Intelligent Machines

In a sense, the entire history of industrial capitalist production has been a history of crowdlabor. The historical process of concentrating working individuals within urban centers, drawing them away from their rural peasant origins by dispossessing those lands from them, is the same process that allowed for the primitive accumulation of capital in Europe during the late middle ages which Marx identifies as an integral prerequisite for the ascendance of modern capitalism. The other major component singled out by Marx in terms of the historical precedents for that ascendence is, of course, early technologies of industrial production such as the steam engine and the novel conventions of rationalized manufacturing which governed them (Marx 1887, 8-11). These two components, the concentration of individuals for collaborative labor and the introduction of machine technologies for industrial production, remain clearly operative as sociotechnical relations within mTurk.

What, then, is the proper distinction between human labor and machine output in the production of ML algorithms? In Volume 1, Chapter 15 of *Capital*, Marx distinguishes between machines and tools on the basis that, while tools may become mechanized to a certain degree, they can only augment the labor-power of the humans who already wield them. Industrial machines, by contrast, are capable of substituting human labor power to a certain degree by imbuing a proportionate amount of their own value into the product (Marx 1887, 275-276). However, this benefit comes necessarily at the cost of an increase in the amount of labor power required from the human worker in their operation of the machine, which does not become profitable for the capitalist unless the rate at which it imbues value into the commodity outpaces this increase. Like all other forms of fixed capital, the computers involved in digital productive activity such as HIT work are not considered agents or conscious participants in the productive process. Yet the Workers who own and operate them do not appear to be considered by their Requesters as meaningfully distinct in this sense from their machines, and are neglected accordingly across the social relations of production, from management to wages to linguistic communications. Through the sociotechnical relations of digital crowdlabor platforms, both the human and the computer are able to be treated essentially as forms of fixed capital – something which would not be possible if not for the specific neoliberal techniques previously described, which ensure that the human worker is no more cognizant of their labor's real use-value than is the computer they use to produce it (Golumbia 2009).

2.5. Critique of Immaterial Inscription

When confronting the extant literature on cognitive or immaterial labor, it would be difficult to justify a lack of direct engagement with the works of Italian Autonomist scholar Maurizio Lazzarato. In "From Capital-Labour to Capital-Life", Lazzarato argues that contemporary capital has developed to a degree of advancement that the production of commodities now produces also entire worlds of social being in the image of the corporation itself via the linguistic process of *inscription,* whereby the logics of the corporate world are conveyed upon the subjects living within them (Lazzarato 2004). Because the contemporary corporation is able to materially construct a real world that encompasses both producer and consumer in their relation to the company, Lazzarato claims, the immaterial logics of that world functionally exist prior to the very concept of the real productive act therein. The general sociotechnical mechanism by which the inscription process is effected – the "expression machine" as Lazzarato calls it – represents for him the major ideological apparatus of contemporary capitalism, exceeding even the means of production, and in doing so reconfiguring the ideal capitalist subject from that of the worker to that of the consumer (Lazzarato 2004, 188-189). If we consider the social relations of mTurk to qualify as one such expression machine inscribed the cultural logics of the DCP upon participating Workers and

Requesters, then it would follow accordingly that all labor contracted therein would be inscribed, prior to the acceptance of any given HIT, via the computational logic which governs the mTurk platform itself.

As the primary vector of linguistic communication between capital and labor within the mTurk environment, the HIT instructions and subjectivizing language of the Turker identity both conform perfectly to the linguistic processes by which Lazzarato claims inscription is effected, with one crucial difference: they are explicitly production-oriented, and "interpellate" their addressees in no uncertain terms as workers, not consumers (Althusser 1971, 171-183). This distinction demarcates the point of departure between this article and Lazzarato in understanding the subjectivities produced within digital crowdlabor platforms, but the aspect of inscription as an explicitly linguistic process occurring effectively prior to all labor activity remains utterly crucial to the proceeding theorization and analyses of this article.

3. Methodology

The critique of immaterial inscription posited in the previous subsection does not dispute the possibility of inscription preceding production, only the presumption that it is exclusive to the realm of consumption. No real relations of consumption are apparent to the Workers of the mTurk DCP. Instead, what are apparent are the reifying relations of production whereby abstract labor is appropriated into standardized labor-power. The inscription process experienced by digital crowdworkers does occur prior to the stage of commodity production, not in consumption but in the initial moment of entry into the DCP labor market – the moment at which a worker becomes inscribed as a Turker, in the case of mTurk. While the particularities of the actual work performed inside the productive stratum of the DCP will naturally vary from HIT to HIT, the semiotic-linguistic landscape of the HIT market which always confronts the worker before they engage in any productive activity at all is homogeneous and universal. Before being accepted, every HIT contract listing is indexed and formatted identically within the page, with the implication being that the labor-power required from one worker for any task which could possibly be contracted through the DCP labor market is in essence equivalent to that of any other.

Therefore, given the fact that Lazzarato's inscription is understood to operate as a semiotic-linguistic process, it cannot be assumed that any manifestation thereof within the textual data of a HIT listing intended for the production of ML would be distinguishable from an HIT listing which is not. For this reason, the methodology of this study will eschew attempts at filtering the available data to include only HITs presumedly intended for ML production, and attempt instead to view the data more holistically.

3.1. The Semiotic-Linguistic Dissolution of the Organic Composition of Digital Capital

Having resolved the question of whether to filter the data corpus based on presumed productive intent, the question arises of how to best index the unfiltered data for analysis. At this point, the concept of *the organic composition of capital* must be introduced. The organic composition of capital is expressed by Karl Marx in Volume 3 of *Capital* by the ratio of c/v where v is variable capital[2] and c is constant capital[3] (Marx 1999). This equation is equal to the difference of the rate of surplus-value extraction[4] over the rate of profit and demonstrates the inverse correlation between the rate of profit[5] and the intensity of constant capital's objectification (Marx 1999, 97-104).

Recalling the conclusion of Subsection 2.4. – that "through the sociotechnical relations of digital crowdlabor platforms, both the human and the computer are able to be treated essentially as forms of fixed capital" – this phenomenon bears additional significance when taken literally in the context of the organic composition of capital, where fixed capital constitutes a definite variable which can be quantitatively measured. This phenomenon then, as a form of inscription per the critique of Lazzarato, is understood as the result of a dissolution of the semiotic-linguistic difference between the workers and the computers which comprise the virtually distributed marketplace of digital crowdlabor, which becomes a dissolution of the organic composition of capital therein as the economic difference between active variable capital and passive constant capital begins to similarly dissolve. In their simultaneous productive consumption into the value-form of digital capital, the abstract labor potential of the human worker is computationally reified as homogeneous labor-power, thereby equating in the final instance the value of the worker's labor-power with the use-value of their computational machines.

The consequences of this equation for the organic composition of capital are hereby hypothesized: the effective price at which labor-power must be purchased is diminished as it becomes conflated with an effective net increase in computational machinic use-value. While the actual use-value of the computational machinery being productively operated must remain mostly constant and only gradually diminishing (as keyboard switches, circuit boards, etcetera wear out with use) there would be no such limitation on the degree to which the humans operating those machines can be computationally inscribed via the increasingly sophisticated computational management techniques deployed by mTurk. Even given such a limitation, the sheer magnitude of additional use-value captured via digital crowdlabor by

---

[2] i.e., the proportion of capital which is ceded by the capitalist to their workers in the form of wages

[3] i.e., the proportion of capital which is employed physically in the productive process and eventually used-up or objectified in this way as part of the product

[4] s/v where s is the surplus-value extracted

[5] $s/(c+v)$

virtually integrating both cognitive labor-power and privately-owned computational machinery into the constant capital variable -- all without having to internalize the accompanying costs -- would still be considerable in terms of potentiated profit.

This potentiation, made possible by the hypothesized semiotic-linguistic dissolution of the organic composition of digital capital within the bounds of the mTurk DCP, represents the methodological lynchpin that binds the political-economic theory of this article to its analytical method. Located at the precise point where the boundary between immaterial linguistic phenomena and material value production becomes porous and permeable, this uniquely alienating relation of production suggests the possibility that a hermeneutic analysis of linguistic data would be capable of yielding insight into the notoriously opaque process of value-formation at the point of production, as is proposed in the following section.

3.2. A Hermeneutic Analysis of the Digital Value-Form

As a work of social science, the intentions and scope of this article should be regarded as the proem to a potentially broader critique of digital political-economy, in the sense that the sociological analysis performed herein serves to collimate otherwise incoherent or incompatible theories of digital capital through the uniquely illuminating lens which happens to be formed by the intersection of digital crowdlabor platforms and machine learning algorithms. In doing so, a lack of sustained engagement may be noted with regards to literature concerning the abstract dynamics of finance and commerce through which the technologies in question are expelled out of the productive circuit of industrial capital circumscribed by the bounds of the mTurk environment, and into that of merchant capital, where they are circulated as realized value-forms. This citational restraint is deliberate and prudent, conforming well to the known principles of the critique of political-economy as an established sociological practice.

In the introduction to his translation of Kojin Karatani's *Marx: Towards the Centre of Possibility*, Gavin Walker cites a passage from Volume 3 of *Capital* in which Marx articulates one such principle by defining the purview of his critical methodology as extending to "only the inner organization of the capitalist mode of production" and no further (Karatani 2020, 19). Walker presents this quote as evidence in his defense of Uno Kōzō, a prominent prewar Japanese Marxist scholar who preceded and greatly influenced Karatani, and whose interpretation of Marx's methodology is widely criticized as being formalistic to the point of esotericism. Walker does not reject the validity of this criticism on its face, countering instead with the assertion that the risks of Uno's method are justifiably taken in the process of abstracting that "inner organization" into a new, superior systematization of the general logic of capital. While the term itself is not invoked in either the frontmatter or translated text of *Towards the Center of Possibility*, what is essentially being proposed by Walker's reading of Uno (and Karatani through him, as will be explored further in later sections of this article) is the theoretical basis for what could be adapted in practice as a methodology of hermeneutic analysis for explicating the value-form.

Such a methodology is proposed herein as a solution to the problem of how to scientifically deconstruct the value-form from within a virtual environment hermetically insulated against any evidence of the commodity-form in its last instance. Within the corpus of data collected for analysis, no information is available other than that which would be immediately apparent to the digital crowdworker at the initial moment of their entry into the mTurk platform. Before being confronted with whatever raw data materials constitute the objects of labor particular to a given HIT, the labor of the crowdworker remains in virtual limbo as they survey the market landscape from above before committing to the contract of a particular HIT listing. Only at that point does abstract labor begin to resemble the concrete labor-power characteristic of industrial capitalism – the same moment at which it disappears beyond the epistemological horizon of the data available to this study. Not coincidentally, this paradoxical situation bears strong resemblance in structure to the paradoxical framework through which Karatani approaches the question of how to objectively conceptualize the value of labor-power prior to its commodification if (as he asserts) the very concept of labor-power as something distinct from abstract labor itself originates from within the commodity-form as something with no real appreciable value prior to its objectification as such (Karatani 2020, 20, 74). By examining Karatani's paradox in the context of the pre-contracted Turker, it should be apparent that these two paradoxes are in fact one and the same, and that they may both be accordingly conceived of in an identical manner.

It has been suggested throughout this article thus far that the human labor-power which is appropriated from the Workers of mTurk is in the broadest possible sense always-already computationalized, or else it would have no use-value within the DCP marketplace. Moreover, it can be further presumed that this computationalization must be a condition independent of labor-power in the concrete, more accurately belonging to abstract labor itself. Otherwise, if the abstract labor of the pre-contracted Turker were not in some sense computational, it could not be appropriated into a form of labor-power with any meaningful value for computational production, and thus could not come into existence in any meaningful sense, per Karatani's paradox. Of course, it must be possible for this transformative valuation between abstract labor and digital labor-power to occur, otherwise there would be no profit to be made on any digital crowdlabor platform at all, which is certainly not the case. Therefore, there must be some means by which both the computationalization of abstract human labor and its valuation as concrete labor-power are accomplished within the pre-contractual HIT listings page of the mTurk platform, and thus accessible for analysis using just the available data corpus.

In Section 4 of Volume 1 of *Capital*, Marx devotes some attention to the question of how exactly it transpires that the infinite potentiality of abstract human labor is reduced down into something standardized and eminently quantifiable in the form of labor-power. Similarly, in Subsection 2.1. of this article, it was briefly described how computationalism is considered by some labor scholars as a managerial strategy for the standardization and agglomeration of labor-power, for the purposes of extracting value therefrom. What Marx concludes regarding this question, is that the *imputation* of value – not the *extraction* of it -- is what accomplishes the homogenization of labor by proxy of its reification within the relations of *exchange*, not of *production*. To translate this hypothesis into the realm of digital crowdlabor, it can thus be said that the real power of the social relations on display within the mTurk HIT market is the fluency with which they "equate as values" (Marx 1887*,* 36) the myriad types of HIT work being contracted, *prior* to the actual acceptance of and productive participation in any given contract.

To understand the means by which digital crowdlabor platforms computationally inscribe and appropriate the labor-power of human workers, it must be understood how that labor-power appears to the workers themselves. This appearance in turn must be understood by deciphering the relation between the objective money-form assumed by variable capital, in its dispensation by the Requester to the Worker, and the real yet intangible value-form of the digital commodities being produced. Within the mTurk DCP, the imputed money-form of variable capital appears transparently to the Turker as the Reward value advertised in each individual HIT listing. These Reward values -- alongside with the username of the Requester, the time allotted for completion, a sentence-length summary of the task, and a few other columns of relevant information – constitute the image of the HIT which is visible most immediately to the Turker prior not only to any engagement in production, but any further virtual navigation of the website whatsoever. It is in this condensed, semiotic-linguistically universal form that the HIT listing appears as the fundamental atomic unit of the DCP marketplace.

By functionally approximating wages, the Reward values thereby appear, to the digital crowdworker who receives them in exchange for their labor-power, as naturally identical to the real value of that labor-power. Bearing in mind Karatani's paradox however, it must be recognized that the real value of any divested labor-power only exists as the expression of a relationship between commodities (Karatani 2020, 74). If, at any moment along the productive circuit of digital capital captured by the mTurk HIT dataset, no real commodities have yet been produced to stand alongside the Reward values which are nevertheless definitely expressed, then it follows that there must exist some other category of commodity somewhere within the system of sociotechnical relations being valuated at that very moment.

This hypothetical class of commodity, it can only be assumed, is the same as that represented by the semiotic-linguistically decomposed constant capital variable in the equation of the organic composition of digital capital that was theorized in Subsection 3.1. If so, then it should be possible to reconstruct the technological character of that variable from just the data available in the mTurk corpus, and in doing so finally obtain a complete systematization of the complex value-form, as it is inscriptively produced within digital crowdlabor platforms. The means of pursuing this reconstructive analysis proposed by this article's methodology is thus: that the listings in the HIT data corpus be reindexed according to the Reward value (as representation of per-HIT variable capital expenditure), and the textual data of the "Title" field[6] be aggregated and linguistically deconstructed via computational content analysis. In this way, it is anticipated that certain linguistic patterns will emerge in conjunction with outliers at the high end of the Reward value index, thereby demystifying the technological composition of the digital capital produced within the mTurk platform.

4. Data and Analysis

The computational content analysis methods of this study were applied to data from a corpus composed of natively digital text collected from the descriptions and publicly visible metadata of active HITs as they appeared on the "All HITs" page, the virtual central square of the mTurk digital labor marketplace (see Appendix A for the web scraper program code). These descriptions and metadata posit an implicit market negotiation between the Requesters and Workers, where the former party attempts to price their HIT so as to minimize their payment to the Workers (and Amazon, which takes a cut) while enticing prospective Workers with the promise of a quick or easy task. There are other factors involved as well, such as selecting for specific 'Qualifications' among the Workers permitted to perform the published HITs given their nature, and the programmed hard-coded time limits for completing the HITs, all of which is at Requesters' discretion.

The corpus of HIT data described above was accumulated over the course of 13 days, from March 9th to March 23rd, 2020 . The corpus consists of 104 documents and the information of 71008 unique HITs in total. The earliest corpus documents are somewhat irregularly paced as evident in their date-time file names, but the majority were extracted, exported, and saved according to an hourly schedule starting at either 12:00 P.M. noon or 1:00 P.M. and iterating from then on for a variable number of hours each day. This iterative technique was necessary for the simple reason that the mTurk HITs page does not store any information on HITs which have been fully completed or expired, which are automatically delisted at that

[6] The "sentence-length summary of the task" for each condensed HIT listing is provided under the "Title" heading.

point. Before being published, every HIT is programmed with a specific finite number of Workers who are able to accept the task before it disappears from the list, and a specific length of time after which it will disappear if it has received no responses. The former number also determines how much money the Requester pays to Amazon up front before the HIT is published. The price of publishing a HIT is computed in total as a function of the number of Workers allowed to accept the task, multiplied by the visibly listed Reward which will be automatically payed-out to the Amazon Payments account of each Worker who has their work approved by the Requester of the HIT, plus a baseline 20% surcharge on that subtotal ("Pricing," accessed February 2021). This surcharge is inclusive of any bonus pay, which Requesters will occasionally offer for exemplary work or as an incentive when they wish to retain the services of certain individual Workers across the duration of multiple HITs.

An additional 5% fee is added to the baseline surcharge for publishing a HIT if the Requester opts to have their HIT listing published with a special limitation that only allows it to be accepted by those Workers who have "demonstrated excellence across a wide range of HITs" and therefor been awarded a "Masters Qualification" by an algorithmic management program within the mTurk platform which "analyzes Worker performance, identifies high performing Workers, and monitors their performance over time," potentially revoking the Qualification at any moment should these Workers fail to meet the opaque expectations of the algorithm's continuous assessments ("FAQs," accessed February 2021). There is also an extensive list of additional small flat fees ($0.50 or less) which apply if the Requester of a HIT choses to restrict their task to Workers with any number of "Premium Qualifications" including many major demographic metrics such as age and income ("Pricing," accessed February 2021). Requesters also have the ability to modify some elements of the default mTurk Qualifications or even design their own entirely, all of which can then be freely used or further modified by other Requesters. For Workers seeking to be granted a specific Qualification, they must file a request form with the original creator of the Qualification, potentially including a "Qualification test" which they must pass for their request to be accepted. In processing the request and determining which ones to accept or reject, the creator of the Qualification is afforded three different methods: an Amazon web service API which algorithmically grades incoming Qualification tests and grants or rejects them automatically, an integral mTurk web service which grades tests procedurally using a grading rubric provided by the Requester and grants or rejects them automatically, or the same integral mTurk web service can be instructed to grant every request automatically with no test required. Alternatively, Workers can be manually assigned Qualifications by Requesters using their unique Worker ID number. Regardless of the method by which it was granted, any Qualification can be revoked at any time for any reason by the Requester who initially granted it ("Creating and Managing Qualifications - Amazon Mechanical Turk," accessed February 2021).

While many of the various categories of data visible from the "All HITs" page were successfully captured and formatted within the full corpus dataset, most of these data will not be discussed further as they did not prove significant within the scope of this study. Specifically, following the methodology proposed in Subsection 3.2., the only data values identified as eminently relevant to the computational content analysis methods employed herein are those included under the Reward and Title headings.

4.1. Full HIT Data Corpus

count	71008.000000
mean	1.136468
std	2.158753
min	0.010000
25%	0.100000
50%	0.400000
75%	1.100000
max	25.000000

Table 1: Statistical Description of Reward Values (Full Corpus)

Before the textual data under the Title headings could be aggregated for analysis, the full data corpus must be reindexed according to the associated Reward values of each unique HIT listing. Once having done so, the statistical distribution of the reindexed Reward data can be described, as in table 1 (see Appendix D for the Full Corpus dataset spreadsheet).

The values shown in table 1[7] (left) describe a statistical distribution where the maximum Reward value is more than 10 standard deviations greater than the mean Reward value, clearly implying the existence of significant outliers towards the upper range of the Reward values in the full data corpus. In order to further explore this implication, the Reward dataset is grouped by value, collapsing HITs associated with identical Reward values into a single datum. In doing so, the grouped dataset acquires the second dimension of "Count", representing the total number of HITs which shared any given Reward value. Having grouped the dataset by Reward and thus expanded it into two dimensions, it becomes possible to visualize the relationship between the Reward and Count variables in the full corpus using a scatter plot, shown in figure 1 (next page).

[7] The "Reward" values in all figures, whether labeled as such or not, are measured in dollars (USD).

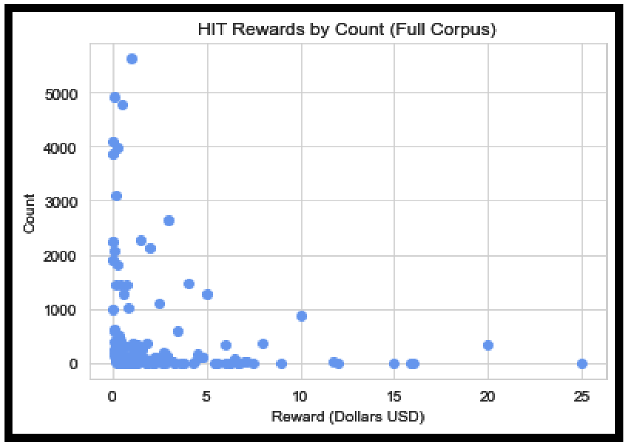

Figure 1: HIT Rewards by Count (Full Corpus)

From the scatter plot in figure 1 (above), it can be seen that there are in fact a relatively small number of outlying HIT listings with unusually great Reward values. Before attempting to draw firm conclusions regarding statistical probability, a more suitable visualization is warranted – specifically, a probability density function (PDF) plot, as in figure 2 (next page).

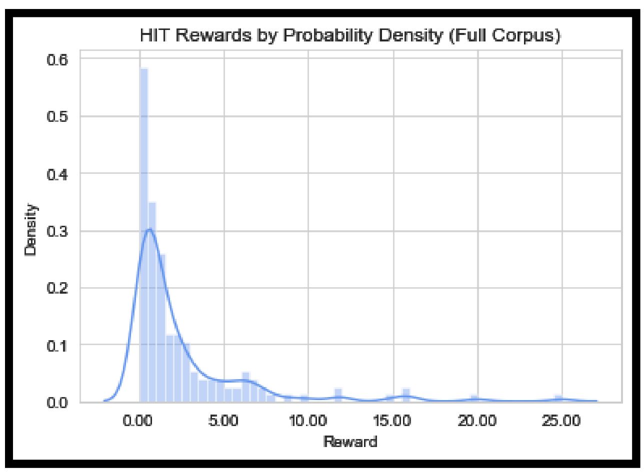

Figure 2: HIT Rewards by Probability Density (Full Corpus)

To produce the plot shown in figure 2[8] (bottom of previous page), the pre-grouped version of the full corpus dataset was used so that the dependent variable represented by the Y-axis in this case is the probability that any individual HIT Reward value will approach the value represented by the X-axis. The figure 2 visualization aggregates the discrete data of the full corpus Reward values into a histogram showing that the vast majority of HITs yield rewards of less than $0.40 and that rewards paying more

[8] (see footnote 8)

than $5.00 were rare. The area under the PDF curve line taken across any given horizontal interval equals the probability of a potential datum falling within the value range represented by that interval. According to the model generated by this function, the probabilities of the Reward values which exceed the X=15 mark are not negative, although they are significantly lower than the rest. The major arc of the PDF curve approaches Y=0 at approximately the X=10 mark, intermittently rising slightly and falling back towards Y=0 at it approaches each cluster of outliers after that point. Judging from this visualization, it would appear more appropriate to consider as outliers any data which exceed the X=10 mark, instead of X=15.

4.2. Top Rewards HIT Data Subcorpus

Based on the interpretation of figures 1 and 2, the outlying data within the full HIT data corpus is defined as those 500 HITs within the horizontal range which is bounded at its upper end by the absolute maximum Reward value (X=25.00) and at its lower end by the value of X=10.00. The data in this range is hereby referred to as the "Top Rewards" data subcorpus and will proceed as the primary object of analysis (see Appendix E for the Top Rewards subcorpus dataset spreadsheet). To better illustrate this boundary as established between the subcorpus and full corpus, a probability density plot is provided in figure 3[9] (next page), where the values for the full HIT data corpus and Top Rewards data subcorpus are shown superimposed over eachother for comparison (see Appendix C for the Full Corpus and Top Rewards subcorpus data visualizations).

[9] The scale of figure 3 has been adjusted to magnify the size of the histogram bars representing the Top Rewards subcorpus comparatively to those representing the full corpus. While the data shown in the plot remain accurate, the visual effects of this ultimately arbitrary manipulation should be noted and taken into consideration as such.

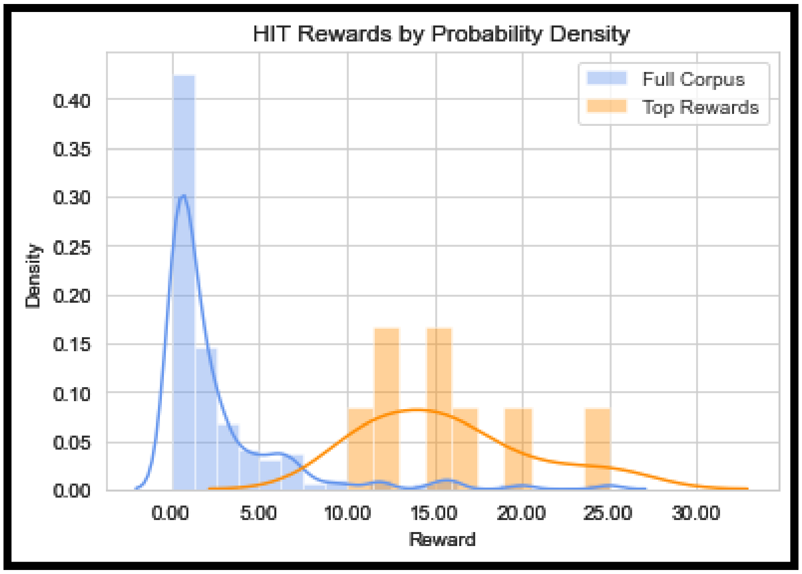

Figure 3: HIT Rewards by Probability Density (Full Corpus and Top Rewards)

The purpose of establishing such an imprecise and relatively arbitrary metric for defining the Top Rewards subcorpus is twofold. First, given that this study's methodology requires the hermeneutic analysis of textual data, the full HIT data corpus must be reduced to a size more amenable to such a method. Second, and more crucially, the ultimate goal of this analysis is to use that data to draw some conclusion regarding the implicit technological composition of the digital capital being produced from the mTurk DCP, and it seems likely that such a conclusion would be most readily drawn from the outlying data at the upper end of the overall Reward distribution. It was theorized in Subsection 3.1. of this article that the production of highly technologized capital should be accompanied by proportionately depressed rates of compensation for the human labor-power involved. Considering then the relatively high degree of technologization to which digital crowdlabor is known to be subjected, the vast majority of the Reward data captured within the mTurk HIT corpus evidently supports this theory by falling well below the standards considered acceptable for non-digital means of labor – except for the outlying data at the upper end of the Reward value spectrum. These outlying Reward values should then, hypothetically, indicate that the HIT work being performed involves a drastically lower degree of technologization than those at the opposite end of the spectrum.

To investigate this hypothesis, the hermeneutic methodology proposed in Subsection 3.2. of this article is applied to the textual "Title" data associated with only those HITs included in the Top Rewards subcorpus for the purpose of discerning whether or not the human labor-power being contracted thereby is indeed being subjected to a drastically lower degree of technologization.

4.3. Computational Content Analysis

The immediate obstacle faced in applying a hermeneutic, semiotic-linguistic analysis to the HIT dataset (even once reduced to just the Top Rewards subcorpus) is the problem of how best to analyze the textual content of the data such that its relevant linguistic trends can be computationally modeled. To accomplish that goal, this study employs an algorithmic Part of Speech (POS) Tagging process (see Appendix B for the POS tagging program code). The POS algorithm classifies any given word within a body of text into its most probable semantic role given the lexical contexts in which it was found – a subjective mode of classification which has been shown to yield more accurate representations of these words' embedded significances than more procedural methods. Once the text of the subcorpus is processed and normalized, and the particular parts of speech are to be analyzed in the subcorpus, the POS algorithm returns for each part of speech a list of words categorized under it and the number of times they appeared within the data. The POS categories available are extensive, but only those which returned unique and intelligible results are included in this study (see Appendix F for the POS analysis data table and data visualizations).

4.3.1. Part of Speech Tagging

The results of the POS tagging analysis reveal the frequencies and overall distribution of the various Nouns, Verbs, Adjectives, Conjunctions/Prepositions, and Foreign Words[10] present in the Top Rewards subcorpus. By illustrating each individual word in terms of its aggregate -- yet nonetheless contextually determined -- lexical function, the POS analysis process perfectly suits the methodological goal of epistemologically normalizing each HIT in relation to each other and deconstructing the entire subcorpus as such. Moreover, by deconstructing the HIT corpus along the dimensions of individual words and their POS categories, insight into the polysemy of those words as linguistic signifiers is gained. It is not sufficient, however, to analyze the deconstructed elements of these data merely in terms of their conventional functions as parts of speech. While those functions do yield insights by isolating the formal mechanisms of language by which broader meanings are thus conveyed to the digital crowdworker

[10]All except one of the non-English words shown in figure 7 were misclassified by the POS algorithm and have been manually sorted under the "Non-English Words" POS column. It is unclear whether these mistakes suggest a general malfunction that may have impacted the accuracy of the POS classifications for the English language words. Regardless, since these classifications proved to be of secondary importance compared to the frequency analysis and manual annotations, this possibility is not investigated further or given significant consideration as a possible source of error.

subject, the ultimate goal for this section of this study remains the investigation of the value semiotics of digital capital, for which there is no readily corresponding computational content analysis technique.

The textual data captured in the HIT corpus constitutes a top-layer slice of the semiotic-linguistic medium through which abstract human labor interfaces with the DCP and, in doing so, is computationally inscribed into digital labor-power. Such a relatively shallow and uncomplex sample of that medium may not necessarily include individual corresponding linguistic signifiers for all social and technological variables involved in the full range of possible interfaces. It can be reasonably presumed, however, that by that same token those variables which are most fundamental to the inscription process would be all the more clearly signified therein. The analysis of this sample is even further streamlined by virtue of being constrained to the Top Rewards subcorpus which, aside from being quantitatively narrower in terms of data to process, is accompanied in the case of this study with a particular interest in the technological composition of digital capital – not an exclusive interest, but a definite direction for analysis with the goal of illuminating otherwise-unaccountable incorporations of fixed capital use-value.

4.3.2. Semiotic-Linguistic Reconstruction of Computational Content Analysis Data

Before the mTurk HIT data can be expected to yield conclusions capable of resolving the theoretical questions so far posed by this study, the Top Rewards subcorpus data linguistically-deconstructed by POS tagging analysis must be semiotically reconstructed along the dimensions of the fundamental variables from which the productive relations of capital itself are composed. These variables are -- in the most basic and reductive sense – the abstract human labor of the worker (“WORKER”), the eminently appropriable homogenized form of labor-power (“LABOR”), and the manifestations of capital held within the relations of production (“CAPITAL”). In this and the following subsections of Section 4 of this article, the typology posited above is practically applied to the results of the POS tagging analysis. Each word returned by the POS tagging algorithm is manually interpreted in terms of which productive relational variable it most apparently signifies and is categorized accordingly. A second, more detailed and descriptive set of categories is developed alongside the aforementioned typology, serving as an additional qualitative dimension for analysis. These categories help to situate individual words within the established relations of production while still preserving a granularity of meaning particular to their conventional significations in language. These two qualitative categories, in combination with the POS tags and aggregated word counts, form the dimensions of analysis visualized in figure 4 (next page).

Part of Speech		Word	Count	Productive Relation	Detailed Interpretation
		DANISH	101	WORKER	Language/Nationality
		Survey	59	LABOR	Data Production (Scientific)
		Compensation	17	LABOR	Management and Arbitration
		screencast	13	CAPITAL	Personal Computational Property
		website	13	CAPITAL	Personal Computational Property
		Business	13	LABOR	Data Production (Social)
		owner	13	LABOR	Data Production (Social)
		laptop	13	CAPITAL	Personal Computational Property
		access	10	CAPITAL	Personal Computational Property
		Thumb	10	OTHER	UNKNOWN
		Language	9	WORKER	Language (General)
		Study	8	LABOR	Data Production (Scientific)
		church	8	LABOR	Data Production (Social)
		sermon	8	LABOR	Data Production (Social)
		video	8	LABOR	Data Production (Visual)
		clip	8	LABOR	Data Production (Visual)
		beta	7	OTHER	UNKNOWN
		transport	7	OTHER	UNKNOWN
	Basic	card	7	OTHER	UNKNOWN
		recordings	324	LABOR	Data Production (Audio)
		speakers	311	WORKER	Language (Spoken)
		minutes	85	WORKER	Management and Arbitration
		SPEAKERS	13	WORKER	Language (Spoken)
		Rules	10	OTHER	UNKNOWN
		data	3	LABOR	Data Production (General)
	Plural	A2HU1HGGYUO3X6	2	WORKER	Management and Arbitration
		webcam	59	CAPITAL	Personal Computational Property
		audio	59	LABOR	Data Production (Audio)
		recording	59	LABOR	Data Production (Audio)
		Social	18	LABOR	Data Production (Social)
		Upload	13	CAPITAL	Personal Computational Property
		Chrome	13	CAPITAL	Personal Computational Property
		Attractiveness	12	LABOR	Data Production (Social)
		Categorization	12	LABOR	Data Production (General)
		Qual	10	WORKER	Management and Arbitration
		Qualification	10	WORKER	Management and Arbitration
		Norms	10	LABOR	Data Production (Social)
		Cognition	9	LABOR	Data Production (Scientific)
		Gambling	8	LABOR	Data Production (Social)
		Online	8	CAPITAL	Personal Computational Property
		Gaming	8	LABOR	Data Production (Social)
		DUTCH	7	WORKER	Language/Nationality
		iPad	3	CAPITAL	Personal Computational Property
		Apple	3	CAPITAL	Personal Computational Property
		Pencil	3	OTHER	UNKNOWN
		Game	2	LABOR	Data Production (Social)
	Proper Singular	Design	2	LABOR	Data Production (Social)
		Interactions	9	LABOR	Data Production (Social)
Noun	Proper Plural	Effects	2	OTHER	UNKNOWN
		Make	324	LABOR	Data Production (General)
		Face	12	LABOR	Data Production (Social)
		Break	10	OTHER	UNKNOWN
		Write	8	LABOR	Data Production (Text)
		say	8	LABOR	Data Production (Audio)
	Basic	Collect	3	LABOR	Data Production (General)
Verb	Present Participle	using	3	OTHER	UNKNOWN
		short	324	WORKER	Management and Arbitration
		Native	311	WORKER	Language/Nationality (Native)
		NORWEGIAN	102	WORKER	Language/Nationality
		Swedish	101	WORKER	Language/Nationality
		NATIVE	13	WORKER	Language/Nationality (Native)
		TAIWANESE	13	WORKER	Language/Nationality
		early	10	OTHER	UNKNOWN
Adjective	Basic	Temporary	1	OTHER	UNKNOWN
		for	45	OTHER	UNKNOWN
		of	23	OTHER	UNKNOWN
		About	8	OTHER	UNKNOWN
		Transcribe	8	LABOR	Data Production (Text)
		during	7	OTHER	UNKNOWN
Conjunction/Preposition		in	1	OTHER	UNKNOWN
		직관적 [Intutive] [Korean]	34	WORKER	Management and Arbitration
		매우높음 [Very high] [Korean]	34	WORKER	Management and Arbitration
		품질 [quality] [Korean]	34	WORKER	Management and Arbitration
		평가 [evaluation] [Korean]	34	WORKER	Management and Arbitration
		소요시간 [Time] [Korean]	34	WORKER	Management and Arbitration
		매우쉬움 [Very easy] [Korean]	34	WORKER	Management and Arbitration
Non-English		人脸吸引力分类 [Facial attract	9	LABOR	Data Production (Social)

Figure 4: POS Analysis Data Table with Word Counts and Categories

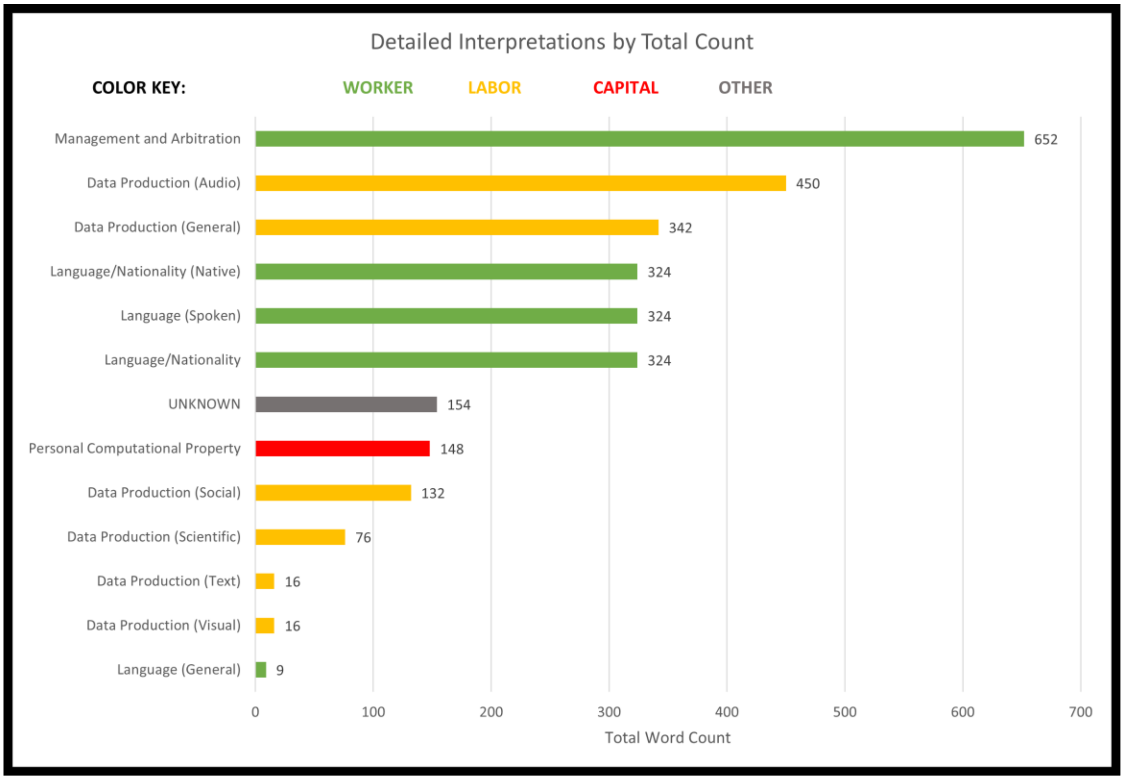

Figure 5: Detailed Interpretations by Total Word Count

The categories shown under the "Detailed Interpretations" column in figure 4 are drawn from the sociological conceptions of digital labor referenced across the various theories cited by this article thus far in addition to the novel theories developed herein regarding computational capital and linguistic semiosis (see figure 5 above for a visualization of these categories and their distributions by word count). It is understood by scholars of digital labor such as Vallas and Aytes that the "fragmentation of tasks" characteristic of digital crowdlabor platforms is accomplished in large part by conventions for strictly curtailing the time allotted for the completion of discreet contracts (Vallas 2019, Aytes 2013). These conventions operate in tandem with the general arbitration of digital crowdlabor, whereby labor-power is subjected to radical exception at the global scale from temporal and political situation alike, allowing for the practically unlimited precision with which Requesters on mTurk are permitted to selectively filter the Workers permitted to engage in production with their HITs via the 'Qualifications' feature of the site. Examining the lists of POS tagging results, the significant frequencies of words such as "short", "Time", "Qual", and "Qualification" are interpreted in this context as signifying the operations of such managerial and arbitrative strategies for regulating abstract labor. Accordingly, these and other such words are categorized under "WORKER" and labeled as "Management and Arbitration" in the detailed interpretation column.

Within the “WORKER” category of words, aside from those deemed indicative of managerial and arbitrative strategies in the general sense, there emerges from the POS analysis results a strong trend of frequently occurring terminology pertinent to language or nationality. These words are either the names of particular languages or nationalities – “DANISH”, “Swedish”, “TAIWANESE”, etc. – or they are general terms directly related to language in general, i.e., “Language” and “speakers”. The majority of the former set of words are categorized by the POS tagging algorithm as adjectives, with a minority being categorized as nouns. Thus, these are taken as most likely possessing a meaning closer to nationality than the corresponding language. This interpretation is corroborated by the observation that, also among the words tagged as adjectives, there is a significant frequency of the word “native” (appearing in both capitalized and all-capitals forms). The cooccurrence of adjectival words signifying specific languages and ones signifying indigeneity suggests that the HITs within the Top Rewards subcorpus are largely targeted towards Workers who are capable of speaking aloud in certain non-English languages, and that the criterion for determining if they are in fact capable enough for the purposes of the HIT work required is in many cases being determined by their identity as a native speaker of that language.

The POS analysis results assigned to the “LABOR” category appear to be highly diverse in terms of both their distribution across the various POS classifications – they are represented in all POS subcategories except for Basic Adjectives and Present Participles. That said, when semiotically reconstructed as signifiers of the HIT work labor-process, the apparent idiosyncrasies within this set of words are superseded in importance by their commonalities as signifiers of digital data production. Even though they span many of the POS subcategories, the majority of the “LABOR” words are tagged as Nouns and Verbs, the former of which all seem to signify particular types of informational objects to be transformed via computational labor-power into data products via the general action words contained within the latter category. The particularities of the informational objects in question are important for their implications for the aforementioned hypothesis regarding HIT work involving spoken language, but equally important is this discovery that the labor-processes for which mTurk Workers’ abstract labor is being computationally inscribed do in fact appear to be formally homogeneous, which supports the more fundamental theory previous cited regarding the necessary conditions for the productive appropriation of human labor by capital.

The third and final productive relation category of “CAPITAL”, under which the fewest words are classified of the three, is represented within only two of the POS subcategories – Basic and Proper Singular Nouns. This conforms quite readily to the basic notion of capital involved in the production-process being *fixed* i.e., materially objectified. What is more interesting about this set of words is that every one of them clearly signifies an object of explicitly computational nature, e.g., a “laptop” or virtual objects such as a “website”, human access of the latter being itself inherently contingent on the materiality of the former. From this observation two conclusions can be drawn: that the only manifestations of capital within the Top Rewards subcorpus which are semiotic-linguistically significant for the inscription of labor-power are explicitly computational in nature, and furthermore that all of these computational manifestations of capital appear to be situated within the possession of the Worker as opposed to that of the Requester. Examining the relatively limited range of words under the “CAPITAL” label in the POS analysis for the Top Rewards subcorpus, there does not appear to be any form of computational technology signified therein which could not be more reasonably interpreted as the centralized productive capital of the Requester than as the decentralized personal property of the Worker. Therefore, all of these words are labeled as “Personal Computational Property” for their detailed interpretation.

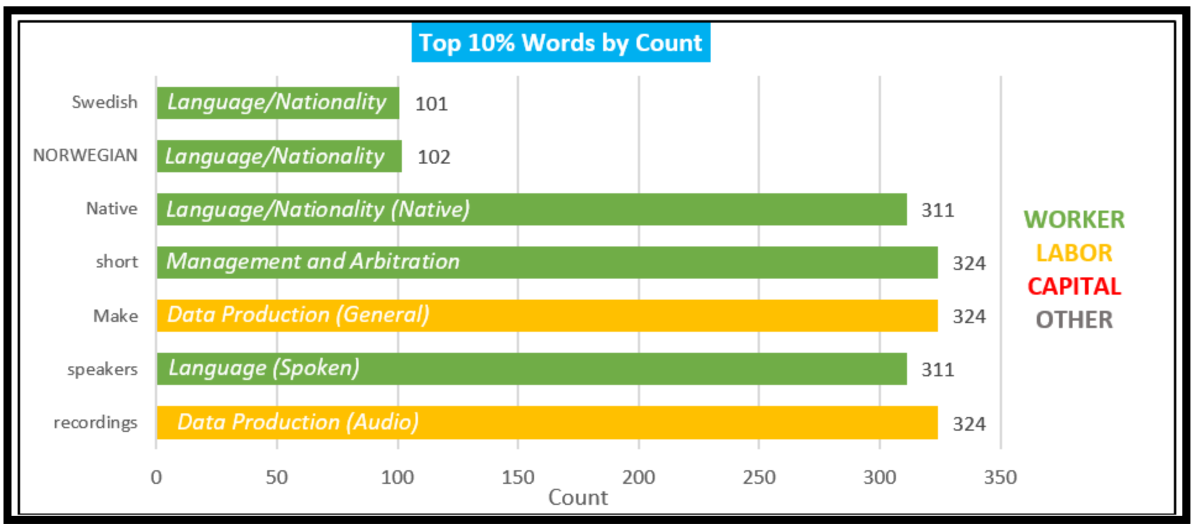

Figure 6: Top 10% Words by Count

The words in figure 4 (page 26) with their Count values highlighted in blue are those which fall within the upper 10% (i.e., 90th percentile) for their individual distributions. These words are isolated for analysis in figure 6 (above) where they are colored according to their Productive Relation categorizations. The majority of the words with the highest individual frequencies are shown to be associated with the “WORKER” category, and the rest associated with the “LABOR” category, with an approximate 2:1 ratio between the total Count values of these two categories (respectively) in their summation. The “OTHER” and “CAPITAL” categories are not featured in this data subset, consistent with the relatively low distributions of those two categories within the greater dataset. What is apparent from figure 6 is that 5 out of the 7 words featured, regardless of prescribed category, are possibly associated with language and/or audio. Even the two outlying words – “short” and “Make” – when considered in this context could be understood as part of instructions for producing language audio data of a particular duration. Thus, it is established by this metric that the 90th percentile of most common words within the Top Rewards subcorpus pertain directly or indirectly to the production of audio language data, constituting the major trend observed therein.

4.4. Analysis of Production and Working Periods in the mTurk Digital Crowdlabor Platform

The linguistic patterns analyzed within the Top Rewards subcorpus are elucidating because of what they illustrate about the particular type of HITs which are most valorized by mTurk Requesters. Through the POS analyses it is shown that the majority of highest-paying HITs on mTurk appear to be designed for producing ML algorithms capable of simulating human speech, and that the labor required of Workers performing these HITs involves interfacing with additional layers of computational technologies in the form of speakers and webcams. Of all the complex and specialized tasks one could imagine mTurk Workers being asked to perform, the type of microlabor shown to posses the greatest amount of value in terms of paid wages seems to involve, in large part, nothing more than a Worker using these technologies to record themselves speaking a certain few languages with native fluency.

Yet, when these HITs are considered as potentially being used for the production of machine learning algorithms, this phenomenon becomes much more economically rationalizable. Specifically, when considering the trends towards recorded language, as opposed to written forms, the requirement that workers be native speakers would make sense if what they are being tasked with involves the sort of verbal fluency and inflection not easily acquired through means other than the natural process of lifelong enculturation. The dispensation of uncharacteristically high Rewards to mTurk Workers (whether through individual HIT listings or in aggregate across many HITs of a particular type) for tasks involving the computational automation of human social performances is easily explicable on the basis that such forms of ML technology remain even today in especially early stages of development as industrially produced computational commodities. As Marx explains in Volume 1 of Capital, the period when machine production is first being introduced into an industry is also the period in which the capital invested into its development will yield the largest profits, which "not only form a source of accelerated [capital] accumulation, but also attract into the favoured sphere of production a large part of the additional social capital that is being constantly created, and is ever on the look-out for new investments" (Marx 1887, 308).

Having thus explained the unusually high floor of the Top Rewards subcorpus HITs, the apparent ceiling observed can be explained as well. Marx, in the section of his 1844 "Economic and Philosophic

Manuscripts" titled "The Struggle Between Capital and Labour and its Results" describes the factors of market economics which determine the absolute upper bound on the prospective valuation of abstract labor for appropriation in the production of any given commodity. This limitation, he explains, is ultimately set according to the absolute conditions of supply and demand – that is, "the demand for labour on the part of capital, and the supply of labour by the working men" (Marx 1844). From this rule, it can be reasonably induced that the apparent upper bound on the value of digitally crowdsourced microlabor is the result of there being a proportionally scarce pool of specialized labor being productively appropriated from native non-English language-speaking Turkers by the mTurk labor market.

Following from this conclusion, the question now arises of how to construe the relationship between the apparent scarcity of the particular Workers whose abstract labor is demanded and the elided use-value of their particular articles of personal computational property which are employed in production alongside them. This relationship -- the relationship between labor and capital in the production of surplus-value -- is understood to be a dialectical one, and as such is not explicable in terms of one element being the primary determinant of the final product. In the case of native language-speaking Workers and computational audio-recording technology, neither is capable of objectifying their latent value into the commodity-form without the other, and should both be analyzed accordingly.

Recalling the detailed interpretations of the "WORKER" category of POS-tagged words in Subsection 4.3.2., there were identified two major trends of semiotic-linguistic signification in the data: those pertaining to languages or nationalities, and those which signified strategies for the management and arbitration of abstract labor. Having already established at this point the theoretical role of the former trend in structuring the production of machine learning algorithms as a digital commodity, the latter trend presents itself for consideration in the service of resolving the question at hand. Of those words interpreted as signifying "Management and Arbitration", those specifically relating to time as a variable are attributed a special importance in the structuring of productive relations within the DCP.

A similar importance is attributed to productive temporality by Marx, who devotes Chapter 12 and 13 in Volume 2 of *Capital* to the development of a theory for explaining precisely how the role of time in the production of surplus-value can be understood. This theory is structured as two layers of nested binary distinction, the first being that between "Working Time"[11] and "Production Time"[12], where the latter always necessarily includes the former, but the former does not always necessarily include the latter

[11] "Working Time" is defined as the time during which labor is engaged in production.
[12] "Production Time" is defined as the time during which capital is engaged in production.

(Marx 1995). The second layer of the theory dissects the category of "Production Time" yet further, resolving it into two distinct *periods* which, unlike the two different *times*, do not overlap:

> In all these cases therefore the production time of the advanced capital consists of two periods: one period during which the capital is engaged in the labour-process and a second period during which its form of existence -- that of an unfinished product -- is abandoned to the sway of natural processes, without being at that time in the labour-process. […] The working period and the production period do not coincide in these cases. *The production period is longer than the working period. But the product is not finished, not ready, hence not fit to be converted from the form of productive into that of commodity-capital until the production period is completed. Consequently, the length of the turnover period increases in proportion to the length of the production time that does not consist of working time* [emphasis added] (Marx 1995, 167-168).

The necessary working period involved in the production of a commodity can be reduced by any number of means depending on the social conventions and material conditions in effect within the industry in question, but all broadly involve either the optimization of productive labor organization or the increased outlay of productive fixed capital. The ultimate purpose, in an economic sense, of reducing the working period for a commodity's production is to accelerate the cycle whereby circulating capital in the form of both labor wages and productively consumed materials is fully objectified into the commodity-form. Only at that point can the complete commodity be entered into circulation as such and the surplus of its total imbued value realized by the capitalist (Marx 1995). Until that point in the cycle, the capitalist must continue to advance capital without return, making the reduction of the working period an imperative tantamount to the primary imperative of capital – circulation itself.

Within the mTurk DCP however, the production-process for each individual HIT contains within itself a complete cycle over the course of which the Worker's personal computational property is appropriated into relations of production as fixed capital, and then is rotated back out into the realm of circulating commodities, where it regains its original status as unproductive property. Because each new iteration of this basic cycle is fully contained within the atomic unit of the HIT contract, there is no point along the cycle itself at which the personal computational property is held as occupying the position of fixed capital while not being actively engaged in the labor-process, and as such there is no effective discontinuity observed between the production period and the working period for digital crowdlabor.

5. Conclusion

Whereas any increase in profits accomplished by the increasing technologization of labor are typically offset by the increasing cost of the productive machinery required in doing so, the productive relations of digital crowdlabor platforms converge along myriad axes – legality, virtuality, even temporality, among many others – to construct a system of exception from that rule. Examining these various axes of exception through the lens of the Amazon Mechanical Turk DCP and its use in the production of machine learning algorithms as a digital commodity has proved to be particularly illuminating, leading to the novel theories posited herein regarding the unique suitability of digital crowdlabor to the historically unprecedented productive mandate now being issued by digital capitalism – the commodification of human cognition itself into the form of artificial intelligence algorithms.

Appendices[13]

Appendix A: Web Scraper Program Code and Documentation

Melamed -- MAPSS THESIS FINAL DRAFT -- Appendix A.zip

Appendix B: Full Corpus POS Tagging and Analysis

Melamed -- MAPSS THESIS FINAL DRAFT

Appendix C: Full Corpus and Top Rewards Subcorpus Data Visualization

Melamed -- MAPSS THESIS FINAL DRAFT

Appendix D: Full Corpus Dataset Spreadsheet

Melamed -- MAPSS THESIS FINAL DRAFT

Appendix E: Top Rewards Subcorpus Dataset Spreadsheet

Melamed -- MAPSS THESIS FINAL DRAFT

Appendix F: POS Analysis Data Table and Data Visualizations

Melamed -- MAPSS FINAL THESIS DRAFT

[13] IF EMBEDDED APPENDIX FILES ARE INACCESSIBLE PLEASE USE THIS LINK FOR ACCESS

Bibliography

"Creating and Managing Qualifications - Amazon Mechanical Turk." Accessed February 11, 2021. https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/Concepts_QualificationsArticle.html#qualifications-and-qualification-requirements.

"FAQs." Amazon Mechanical Turk. Amazon Mechanical Turk, Inc. Accessed February 11, 2021. https://www.mturk.com/help.

"Pricing." Amazon Mechanical Turk. Amazon Mechanical Turk, Inc. Accessed February 11, 2021. https://www.mturk.com/pricing.

Althusser, Louis. "Ideology and the State." Lenin and Philosophy and Other Essays 2 (1971).

Aytes, Aythan. "Return of the Crowds: Mechanical Turk and Neoliberal States of Exception." In Digital Labor: The Internet as Playground and Factory, edited by Trebor Scholtz. New York, NY: Routledge, 2013.

Callison-Burch, Chris, and Mark Dredze. "Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk." In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010.

Dyer-Witheford, Nick, Atle Mikkola Kjøsen, and James Steinhoff. "Inhuman Power." Artificial Intelligence and the Future of Capitalism. London: Pluto Press, 2019.

Ellmer, Markus. "The Digital Division of Labor: Socially Constructed Design Patterns of Amazon Mechanical Turk and the Governing of Human Computation Labor," 2015, 14.

Felstiner, Alek. "Working the Crowd: Employment and Labor Law in the Crowdsourcing Industry." Berkeley Journal of Employment and Labor Law 32, no. 1 (2011): 143–203.

Golumbia, David. The Cultural Logic of Computation. Harvard University Press, 2009.

Gupta, Neha, David Martin, Benjamin V Hanrahan, and Jacki O'Neill. "Turk-Life in India." In Proceedings of the 18th International Conference on Supporting Group Work, 1–11, 2014.

Hara, Kotaro, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. "Tohme: Detecting Curb Ramps in Google Street View Using Crowdsourcing, Computer Vision, and Machine Learning." In Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, 189–204, 2014.

Hardt, M., and Antonio Negri. Declaration. Argo-Navis Author Services, 2012.

Irani, Lilly C, and M Six Silberman. "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 611–20, 2013.

Irani, Lilly. "Difference and Dependence among Digital Workers: The Case of Amazon Mechanical Turk." South Atlantic Quarterly 114, no. 1 (January 1, 2015): 225–34.

Karatani, K. Marx: Towards the Centre of Possibility. Edited by G. Walker. Verso Books, 2020.

Laws, Florian, Christian Scheible, and Hinrich Schütze. "Active Learning with Amazon Mechanical Turk." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1546–56, 2011.

Lawson, Nolan, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. "Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk." In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 71–79, 2010.

Lazzarato, Maurizio. "From Capital-Labour to Capital-Life." Ephemera 4, no. 3 (2004): 187–208.

Lazzarato, Maurizio. Contemporary Marxist Theory: A Reader. Edited by Andrew Pendakis. New York: Bloomsbury Academic, an imprint of Bloomsbury Publishing Inc, 2014.

Le, John, Andy Edmonds, Vaughn Hester, and Lukas Biewald. "Ensuring Quality in Crowdsourced Search Relevance Evaluation: The Effects of Training Question Distribution." In SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, 2126:22–32, 2010.

Marx, Karl. "Capital Volume III." Economic Manuscripts: Capital, Vol.3, Table of Contents. International Publishers, NY, [n.d.], 1999. https://www.marxists.org/archive/marx/works/1894-c3/.

Marx, Karl. Capital: Volume 1. Translated by Samuel Moore and Edward Aveling. First english edition of 1887. Vol. 1. 3 vols. Critique of Political-economy, 1887.

Marx, Karl. Grundrisse: Foundations of the Critique of Political-economy. Translated by Nicolaus Martin. Penguin UK, 1973.

Marx, Karl. Translated by Martin Milligan. Marx's Economic and Philosophic Manuscripts of 1844. Progress Publishers, Moscow 1959; 2000. https://www.marxists.org/archive/marx/works/1844/manuscripts/preface.htm.

Mitra, Tanushree, Clayton J Hutto, and Eric Gilbert. "Comparing Person-and Process-Centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk." In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 1345–54, 2015.

Quinn, Alexander J, Benjamin B Bederson, Tom Yeh, and Jimmy Lin. "Crowdflow: Integrating Machine Learning with Mechanical Turk for Speed-Cost-Quality Flexibility." Better Performance over Iterations, 2010.

Reynolds, Kelly, April Kontostathis, and Lynne Edwards. "Using Machine Learning to Detect Cyberbullying." In 2011 10th International Conference on Machine Learning and Applications and Workshops, 2:241–44. IEEE, 2011.

Slater, Don, and Andrew Barry. Technological Economy. Routledge, 2005.

Sriraman, Anand, Jonathan Bragg, and Anand Kulkarni. "Worker-Owned Cooperative Models for Training Artificial Intelligence." In Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17 Companion, 311–14. Portland, Oregon, USA: ACM Press, 2017.

Vallas, Steven P. "Platform Capitalism: What's at Stake for Workers?," 2019, 13.

Zhao, Liyue, Gita Sukthankar, and Rahul Sukthankar. "Incremental Relabeling for Active Learning with Noisy Crowdsourced Annotations." In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 728–33. IEEE, 2011.