ECONOMY OF THE DATASET: MARX AND LARGE LANGUAGE MODELS

ANDRE YE

An analysis shows that the massive datasets which are enclosed en masse and ingested by gargantuan language models are, in reality, very queer things, abounding in metaphysical subtleties and theological niceties. So far as they are values in use, there is nothing mysterious about them, whether we consider them from the point of view that by their properties they are capable of satisfying steps towards the goals of autonomous 'intelligence' development, or from the point that those properties are the product of human labor. It is as clear as noon-day, that the data lassoer, by their industry, changes the forms of the unstructured data furnished by the collective Internet (public and private), in such a way as to make these useful to them. The form of content on a webpage, for instance, is altered, by making several dataset samples out of it. Yet, for all that, these dataset samples continue to be that common, every-day thing – another page of content. But, so soon as it steps forth as an element of a Dataset, it is changed into something transcendent. It not only stands with its body on the ground, but, in relation to all other Dataset-elements, it stands on its skin, and evolves out of its digital brain grotesque ideas, far more wonderful than "data collection" ever was.

My only justification for shamelessly plagiarizing Marx is that Marx has set forth an interesting and important framework for thinking about how today's breeds of large language models as varied as ChatGPT, Delphi, and BLOOM consume and manipulate information to ultimately give rise to an *economy of the Dataset*: a system of information-exchange in continual circulation and agitation. I will argue that Marx's concepts of the commodity, commodity fetishism, and capital allow us to gain an important philosophical dimension on understanding dataset samples, labor and representation in datasets, and the question of 'new' or 'original' intelligent production, respectively. In doing so, I have three different goals for three different audiences: for the computer scientist, to develop some philosophical consciousness 'outside of' their technical episteme; for the layperson, to resist or at least complicate simple explanations by computer scientists on how language models work; and for the philosopher, to advance a set of technically informed modes of problematizing and investigating these new technologies.

It should strike as at least somewhat strange that we would think to group together information from sources as varied as Elon Musk's tweet history, a Wikipedia page on Barthes' "death of the author", and this CNN guide to Brexit. Proto-datasets for specific-purpose models were gathered as arrangements of data with particular *content* – content being the set of attributes from which meaning arises. For instance, one collects English and French text pairs as a prerequisite to building an English-to-French and French-to-English translation model, or one collects question-answer text pairs as a prerequisite for building a question answering model. Such proto-datasets included only data which satisfied particular content requirements, both within each sample and across samples. But it is with the dominance of general-purpose models that the true significance of the Dataset has emerged: a grouping together of data not by content, but rather solely by *form* – the body through which content is given its existence. For general-purpose language models, this form must be text. Here, I mean text both in the philosophical and in the technical sense: text as an irreducible complex web of language, ideas, and cultural references that is constantly in flux (qua Derrida), sure, but equally so the system of encodings, conventions, and technological apparatuses which allow me to type, send, read characters on a screen. Form is, to be reductive, the precondition for content. It is through the idea of a Dataset as a collection of forms, departing from previous restriction by content, that we can understand its "metaphysical subtleties and theological niceties".

It is precisely this point which allows us to begin establishing parallels between Marx's analysis of the commodity and to theorize an economy of the Dataset, by beginning with *form as money*. To begin with, it should be noted that my notion of the economy is not strictly concerned with the operation of exchanging goods which is so fundamental to the traditional study of political economy. I do not mean to suggest that three Wikipedia pages on "Irony" is equal to ten dataset-dollars, for which I can buy one Reboot subscription page: this is too literal of a reading. I am more interested in an economy of the Dataset in the sense that it becomes possible to think of and treat elements of general datasets as relating to each other as commodities – as being comparable, quantifiable, and exchangeable not between individuals but in their inclusion in said datasets. Just as an economy of goods begins with money, an economy of the Dataset begins with form. Marx writes in the beginnings of *Capital* that money must "play within the world of commodities the part of the universal equivalent": it is the glue which makes it possible to link together all commodities in a web of hypothetical exchange. Form plays the "special social function" of money: it allows the computer scientist to see within every webpage its form, and therefore its potential for inclusion within the Dataset; in the same way that I see within every fruit in the produce aisle its price, and therefore its potential to be purchased for by me. Such a Dataset is much more a 'world of commodities' than proto-datasets: I do not see Риа Новости as potential for annexation into a proto-dataset for French-English translation (I cannot commodify it), in the same way that I do not (or should not) see the people walking along the street as potentials for annexation into my romantic life (although romance is arguably already a thoroughly commodified category). The logic of forms, then, is the logic of money, and this is the precondition for how the computer scientist approaches and sees data.

This elevation of form becomes intuitive, then natural, for the computer scientist. It is no surprise that dominant research subareas in deep learning are stratified along form: computer vision for data in the form of images, signal processing for data in the form of audio, language modeling for data in the form of text, and so on. It is worth pondering why deep learning is not principally stratified by content: ethical judgment modeling, conversational modeling, future-forecasting, creative modeling, and so on. The argument that the stratification by form is technical is not entirely true nor fair, because now there is heavy technical sharing across these subareas (a lot of attention, it turns out, really is all you need). Although there is an increasing amount of multimodal research which blurs the lines between these stratifications, this stratification still represents the set of terms which computer scientists dominantly use to think about data: 'this is text', 'that is an image', 'that is audio' – less often fundamentally as as 'this is an ethical judgment', 'that is a natural world representation', and so on. It is worth noting that the science-fiction imaginary of the 1900s tended to think of artificial intelligences along lines of content rather than form: HAL 9000 from *2001: A Space Odyssey* is built for advisory and conversational purposes, but not for ethical judgements – and it is precisely the tension between these two which provides the basis for the plot (an AI on a mission without a moral compass – what will happen?). This now-dated conception did not consider division along forms to be the primary problem to confront: they took as a given the capability of synthesis across forms, which is today still an ongoing research problem in deep learning. While writing this essay, I have often myself using the form of data to refer to the data, such as saying 'the model was trained on text' or 'these images…', and I indeed cannot think of what other name I would use. This is an example of the schizophrenic force of synecdoche, in which a part is used to indicate the whole, such as giving 'your hand' in marriage to indicate giving your body and life or feeding hungry mouths to indicate nourishing bodies. Synecdoches are interesting because it is never just that the part represents the whole, but that the part subtly becomes the whole as we know it. We become obsessed with parts as if they were the wholes, and over time we fail to remember that parts were attached to wholes to begin with. This is the hyperreality which French philosopher and cultural theorist Jean Baudrillard wrote of. The computer scientist, too, begins to see the commodity in every new article, comment, video – the potential for inclusion into the dataset –they begin to see information as its form, and moreover treat it as its form: this is the fetishization of the commodity.

What happens when the form of data becomes its central precondition for considering its relation to other data in the Dataset – when its form becomes it? Marx shows that subjects in the world of commodities partake in commodity fetishism, in which the necessary abstraction of commodities into terms of their equivalence and exchange against other commodities (i.e. by money or form) allows for them to become essentialized by such terms, and for these terms to be perceived as 'real' and 'authentic' metaphysical identities. These are the 'grotesque ideas' which emerge from data's 'digital brain'. Most importantly, the fetishism of the commodity obscures the real history of labor which has made the commodity into what it presently is. I see that the iPhone 14 is $799 at the time of writing, but not the processes by which workers put their labor into its production, in this case in highly exploitative contexts. I see this student opinion article on tourism and imperialism as text, and therefore as a commodity suitable for annexation into the Dataset, but not the intellectual labor and thinking which the author has put into making its content. Marx writes of commodity fetishism:

> ***A commodity is therefore a mysterious thing, simply because in it the social character of men's labour appears to them as an objective character stamped upon the product of that labour; because the relation of the producers to the sum total of their own labour is presented to them as a social relation, existing not between themselves, but between the products of their labour. This is the reason why the products of labour become commodities, social things whose qualities are at the same time perceptible and imperceptible by the senses.***

Bringing products of labor into the socialized world of commodities requires an obfuscation, or even a shedding, of the history of labor. The computer scientist's construction of massive unstructured datasets requires this very obfuscation: in the process of scraping data from webpages, text becomes coldly excised from its original conceptual contextualization and placed within the sterile laboratory environment of the Dataset. It is now another commodity neatly tucked into a large silo, surrounded by other commodities surgically cloned from opposite ends of the Internet. What better real-life analog is there to a Dataset than produce supermarkets, where bananas from Ecuador are placed next to apples from Washington and mangos from the Philippines, and large plastic signs display their relative prices so I may choose between them as commodities (the logic of the consumer)?

Nowhere is this fetishism more pressingly apparent than in the way in which the relation between language models and datasets is described, both for popular and technical audiences. It is often said that models are 'trained on large amounts of data', 'read large swathes of the Internet', 'learn to write from lots of text'. A more technically accurate description might even be that models 'learn to predict the next token (word or word-part) given a previous sequence of tokens'. These statements are philosophically incorrect, in the sense that they participate (unwittingly) in the commodity fetishism of the Dataset. The data in the Dataset becomes a very queer thing: it becomes an autonomous metaphysical object, somehow a thing for itself, which is given to the model. What is obscured is the production of the data by humans. The philosophically correct formulation which recognizes the formative labor history of dataset elements is 'the model learns to imitate how a subject represents an idea, across all subjects which have contributed to the dataset'. It is not that this philosophically correct formulation is 'secretly' known, but that we favor the philosophically incorrect formulation for brevity or convenience: it is precisely this instinctive favoring for convenience which is the fetishism.

In this way, the claims that software-writing language models or text-to-image generation models unfairly 'stole' and failed to give credit to the writers, coders, artists, and other individuals should first look towards this fetishization of information – a fetishization which happens even before there is any model at all, and a fetishization which seems quite innocuous as protected by legal apparatuses. Perhaps it is time for us to reconsider the origin of our problematizations.

There is one more question – so what is the role of the model in the Dataset? The function of the model, I hypothesize – and this is, admittedly, a weak hypothesis – is analogous to the production of capital under a Marxian analysis of capitalist economy. The basic formula of commodity exchange is A – M – B, through which a commodity A is sold for some quantity of money M, which is used to purchase another commodity B. The basic formula of capital's growth, however, is C – A – C', through which principal capital C is invested into commodity A and sold for a higher quantity of money C'. This formula is an ultimate fetishism: the commodity fails to be an end through which money can help us obtain, but rather the commodity is a tool through which capital embodies itself in its eternal quest for expansion and accumulation. The passage from simple commodity exchange to capital accumulation is parallel to the passage from narrow language modeling to general language modeling. A French-to-English translation model maps a French text to its fundamental 'meaning', which is exchanged for an English body. But general language models are not confined to a particular domain of meaning: their domain is the Dataset, an open world of commodities. Large language models must generate new forms from existing forms, new forms which circulate and become reinvested in the process of creating new forms. Such models partake in this novel economy, this economy of the Dataset. The role of the language model in this economy is to instigate circulation, and to produce surplus capital from this circulation. The Dataset itself is a sterile and structured container, and the relations of comparison and exchange exist conceptually. But a model actualizes these relations materially: items in the dataset are brought in and out of each other, repeatedly pushed into and out of the model, partially imprinted upon its surface.

Large language models are shiny, glitzy cultural myths. They are attractive interfaces which possess both material engagement and the bizarre transcendental dimension of magic. But what is more interesting for me is its premise – the economy of the Dataset – a premise which is too often overlooked, seen without being seen. If our critical analyses begin at the model, then in a certain way we have already accepted its precondition, the fetishization of the Dataset. Even as we find ourselves in a purported age of blindingly fast technological development, technologies are always imbued by the structures that provide the conditions of their production. We should not depoliticize large language models as posing broad abstract, humanitarian threats: they are specifically situated, and a productive analysis must begin from these situated knowledges – these very *queer* ways of knowing and seeing the world.