Data Diets and Democracy: The Chinese Social Credit System From a Machine Learning Perspective

Joshua Fairfield

2019-06-25T09:30:57

Introduction

The current debate asks whether the set of technologies that make up the Chinese Social Credit System (CSCS) trends toward digital dictatorship or digital republic. I make two points. First, Chinese AI may well become stronger than Western AI because of differences in the data available to train machine learning algorithms. Second, the mechanisms of machine learning may tend toward a reversal of the longstanding advantages of information decentralization in capitalist economies, with negative knock-on effects for liberal democracies worldwide.

My conclusions: the CSCS trends against democracy. It is being built by a competent and motivated anti-democratic system with social control as one stated goal. The more important question though is whether the Chinese machine learning data diet will make Chinese AI stronger than Western AI, and whether the realities of machine learning will undermine Western-style capitalism and liberal democracy. As this essay argues, I think there is a real chance that both will occur.

These conclusions place the current piece in both agreement with and at odds with the framing arguments for this debate. Both Reijers and van ‘t Klooster move from talking about *the* CSCS to a discussion of *a* SCS. That move is correct in that the surface question of whether the Chinese version of these systems will be turned to anti-democratic ends is not particularly interesting. But machine learning algorithms cannot be divorced from the data they study – data produced by citizens in a given context. Thus, I place the machine learning algorithms back into context by talking about the role that citizen-produced data plays in training machine learning algorithms. For three reasons, discussed below, Chinese machine learning algorithms may benefit from more and more varied data than do Western versions, and this in turn may impact the underpinnings of capitalism and democracy worldwide.

The Chinese Data Diet

The key to machine learning is data, its volume, completeness, and variety. Recent advances in algorithm design have been merely (or mostly) advances in the amount, variety, and quality of the data we have been able to feed machine learning algorithms. Thus, to the extent that the CSCS represents something different from the emerging Western system of artificial-intelligence-driven pricing (of credit as of any other commodity or service), the differences will need to lie in the data channelled into these systems, and the differences that the artificial intelligences will learn on a different data diet.

I then leave the reader with some musing about what it would mean if the CSCS and similar systems generate superior machine learning results because of improved access to datasets, amount of data, and the new on-ramps created by the physical systems of the CSCS (neighbour reporting and the like). In particular, I suggest a dangerous potential overturning of a longstanding principle of capitalist development. Capitalism in its late-20th century idiom has succeeded largely because decentralised information processing is far more efficient than is centralised data processing. The 'Invisible Hand' of the market is a metaphor for the coordinating actions of decentralised economic decision makers, who organise production around price points, for example. The question is whether the strong synergies of centralisation inherent in machine learning will undermine this economic consensus, and thus undermine liberal democracies in general.

Machine learning and related technologies consist of a range of techniques, the full scope of which is outside the reach of this article. One shift, however, is worth mentioning. Many of the algorithms we rely upon have existed with minor updates for a number of decades. What changed in the early 21st century is that the raw volume of data available to train machine learning algorithms exploded.

Data matters because, for at least many of the most common and most effective machine learning programs, the machine does not understand anything. Rather, it finds connections in data by looking at a lot of it. Machine learning algorithms rely on correlations in the data that are relevant to the prediction the algorithm is trying to make. The more data a machine learning algorithm can ingest, the more correlations it can find. If one's concern is pure accuracy and not explainability, the weights used by a machine learning program can be quite extensive: perhaps my political reliability or creditworthiness correlates with my social media contacts, recent purchases, books read, social media commentary, outcomes of judicial processes, and so on. Each correlation may yield information gain, which will help in making a better selection.

The difference in the machine learning algorithms created for advertisement, election influence, and creditworthiness in the West and those that result from the CSCS – *if* there are important and salient differences – can be usefully explained in terms of the differences in data diet that the CSCS system affords. This point has gone underexplored in the debate. The following parts discuss how the CSCS might differ from the Western machine learning context in terms of access to data, amount of data, and physical on-ramping.

Access to Data

Data in the United States and Europe is often kept separate from other data pools. This is often for reasons more of competition than privacy. Facebook's knowledge of its users is its goldmine; likewise, Google's tracking of users as they interact with websites across the internet. A machine learning algorithm that had access to the combined Facebook and Google datasets would be able to extract out features and would undoubtedly be more accurate than one based on either browser-tracking or social media interactions alone. Unintentionally, this offers some protection against machine learning algorithms. The question is whether the Chinese political structure offers avenues for combining datasets that western countries do not.

First, the Chinese government has a stronger hand in the information industry than do Western governments. This government whip-hand sometimes merely increases raw government power, for example, in the form of country-wide censorship mandated by the government and carried out by industry. Second, the Chinese government's overt encouragement of the CSCS overlaps with its public commitment and large investment in artificial intelligence. In addition, powerful consortia vie for the political and commercial capital at play. My view is that there is ample cause to believe that these political considerations may cause Chinese AI to benefit from larger and more varied datasets than do those trained by individual U.S. companies, no matter how large.

Amount of Data

A key question is whether the Chinese legal, cultural, and technological frameworks permit access to more, different, and more granular data. More data is easy to imagine, but hard to confirm. The population numbers are higher: some estimates put the number of mobile internet users in China at more than double the entire United States population. That is a lot of smartphones recording a lot of data, which can be used for targeted behavioural advertising, political verification, and so forth.

However, overlapping data is often less useful than different datasets. Data about you as you surf the web is even more useful when combined with data from your smartphone geolocator, or your smart television's parsing of your conversations and viewing habits. This is called sensor fusion: combining multiple datasets often results in information beyond their sum. Sensor fusion has a greater impact in deeply gadgeted societies. I am aware of no study on comparative Western-Chinese sensor fusion. Such a study would prove deeply informative on the relative strengths and weaknesses of Chinese machine learning.

Municipal On-Ramping

The last element is whether the Chinese government support, especially in the form of model city initiatives and the structure of bureaucratic advancement within the Chinese Communist Party, has a strong impact on building out physical sensors, like smart cities or municipal sensors, that serve to on-ramp citizens' data and make it available for machine learning.

I am not certain that government power and different legal and social norms surrounding privacy make much of a difference for physical on-ramping, at least when one compares China to the United States and the U.K. In those countries, the temptation to circumvent constitutional restrictions on government mass surveillance by routing the data collection through third-party firms has proven irresistible. In continental Europe, norms surrounding personal privacy in public spaces take on a different aspect, and we might expect AI trained on sparser European data to be at some disadvantage.

Physical on-ramping, in the form of municipal cameras, pressure sensors, and other so-called “smart city” technology, seems to be developing similar amounts of coverage in China and the United States. The mechanisms may be slightly more commercially driven in the U.S., and the government plays more of a recipient and beneficiary of commercial datasets, which it turns to policing and intelligence purposes, but I have no sense that the amount and flow of data across physical sensor networks is less.

My understanding is that often party advancement is tied to furtherance of party buzz words and goals. To the extent Social Credit is a party buzz word, we may expect municipal and regional officials to make efforts to develop facilities for the CSCS in their areas of expertise and authority. If, then, party members who are in positions of municipal authority believe that physical on-ramping of municipal sensor data is valuable to their political careers – as I think evidence indicates is the case at least at present – then we might expect data on-ramping through municipal sensors to strongly exceed at least European rates.

Concluding Thoughts

I have raised these points because they are salient to the two fundamental questions posed by the CSCS. Those questions are political and economic, if one sees those as separate spheres. The political element, which has attracted the most attention, is the extent to which the CSCS will be profoundly anti-democratic. Since the CSCS was designed by and for a non-democratic political system, this is a bit of a non-question. Of course, the system is being used to enforce political hygiene, and there is already more than enough evidence of these technologies being used to monitor and oppress.

The follow-on problem is harder. AI in general, and machine learning as its most currently successful instantiation, may destabilise the current consensus in economics that decentralised information processes in markets, coordinated primarily by price points, are the best way to run an economic system. It is not inconceivable that machine learning may change the playing field of capitalism. I do not suggest that centralised AI will be able to guide a national economy better than will some form of combined machine learning and decentralised firm-based decision-making, but I do suggest that the balance of those contributing elements may shift decidedly.