STANDARDIZED DATA COLLECTION: LEGAL REQUIREMENTS, GUIDELINES, OR COMPETITION?

- *Frank Fagan**

I. INTRODUCTION

Imagine that a U.S. bank wishes to develop a predictive model for granting credit to borrowers below the poverty line. With the current U.S. population at 325 million—of which nearly 40 million live in poverty[140]—the universe of available data for making predictions may be limited. Even if a quarter seek loans, and the bank has experience lending to 10% (or 4 million borrowers), the amount of predictive precision required for avoiding bad loans and creating a profitable lending business may be insufficient nonetheless since data about past loans may be inadequate.[141] In other national markets, where the population is larger—in particular the number of those living in poverty—data may be available for developing a sufficiently precise predictive model.[142] That model would obviously be profitable in its country of origin, but for export, the predictive patterns that it identifies at home must also be present abroad. In terms of industrial strategy, developers of predictive models for export could collect and test general stocks

* Associate Professor of Law, EDHEC Business School, France. I thank Dean Ranita Nagar for her invitation to submit this Essay to the *GNLU Journal of Law & Economics*, and for comments, Saul Levmore.

[140] *Basic Statistics*, TALK POVERTY, http://www.talkpoverty.org/basics (last visited Feb. 24, 2019).

[141] Note that, with current technology, "supervised deep learning algorithm will generally achieve acceptable performance with around 5,000 labeled examples per category and will match or exceed human performance when trained with a dataset containing at least 10 million labeled examples." IAN GOODFELLOW ET AL., DEEP LEARNING 2 (2016). Thus, the bank may have insufficient data even if it can purchase other data from data brokers, especially in contexts where counterfactuals matter, but remain generally unobservable. In this example, the bank may be restricted by profit margins from observing the outcome of granting loans to those whom the model borderline rejects. It might "invest" in developing a more precise predictive model by randomly granting loans to the rejected, losing some money in the process, and then teaching the model from those random loan observations to enhance decision making accuracy in the future. But this method will reduce current lending margins and may not be profitable in a present value sense in some markets, inhibiting a project from taking place. See Frank Fagan & Saul Levmore, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CAL. L. REV. *9-10 (forthcoming 2019), which discusses the same problem within the context of unobserved flight of arrestees who are denied bail. Instead of negative net present value, the problem with randomly granting bail in the arrestee example is potential equal protection violations and arbitrariness through random application of rules. On equal protection violations, see Michael Abramowicz, Ian Ayres & Yair Listokin, *Randomizing Law*, 159 U. PENN. L. REV. 929, 964-74 (2011). For a discussion of problems with arbitrariness, see RONALD DWORKIN, LAW'S EMPIRE 178-84 (1986).

[142] Of course one million observations may be sufficient for developing a useful and profitable predictive model. *See* GOODFELLOW ET AL., *id.* at 141. The example merely demonstrates the intuition of the problem. So long as generalizable data is collected and mined at a lower cost in the exporting country, there exists an opportunity for predictive model exporting.

ISSN 2582-2667

of data alongside country-specific ones. In terms of policy, law could nurture low-cost data collection that stimulates the construction of models at home.

But law could go a step further and additionally encourage the development of broadly useful predictive models, especially in national machine-learning-based infrastructure investments. This can be done with substantive data collection requirements in exchange for government funding or tax incentives, or the development and announcement of process-based standards for data collection.[143] Imposing substantive data collection requirements in exchange for funding is efficient inasmuch as the project is beneficial and the additional requirements can be profitably used in other contexts. The imposition of process-based standards entails social cost, but the provision of guidelines may be enough to reap the rewards of standardization when the private benefits from data independence are small. Efficient standards may fail to emerge, however, even with law's endorsement, in the presence of severe collective action problems.[144] Of course, the danger of endorsement is that the standard itself is inefficient. Competition among jurisdictions—in particular, a national desire to win the global AI race—may be expected to bring about efficient results, but only if big data is big enough within jurisdictions or across the jurisdictions of federated partners.

All of this is consistent with the problem (and general mystery) of choosing between the benefits of competition and economies of scale. Technical data collection standards present the added complexity that lawmakers may be unable to distinguish between efficient and inefficient leapfrogging. In other words, do the presumed economies enabled by standards today outweigh the drag on the potentially beneficial standards of tomorrow? And will mandating standards today eliminate the possibility that future and superior standards will arise? The answers to these questions are perhaps, at this point, still irregular enough to be empirical, and in any case, are left for future work. Today, on

[143] Standardized data collection furthers data portability and interoperability, which are often pre-conditions for cross-firm and cross-industry data exchange. *See* Pol'y Dep't A: Econ. & Sci. Pol'y, Eur. Parl., *Industry 4.0* (Feb. 2016), http://www.europarl.europa.eu/RegData/etudes/STUD/2016/570007/IPOL_STU(2016)570007_EN.pdf; *see also* FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY 2 (2014) (on consumer data collection practices).

[144] *See infra* § III.A.

the other hand, surely the benefits of standardization must be discounted by an uncertain future. Standards may generate economies of scale, but they simultaneously inhibit competition and its benefits. This is the danger of centralized standards either imposed or announced. Good arguments for economies of scale can easily be made but difficult to believe upon further scrutiny. In other settings, auctions can serve as scrutinizers, though perhaps here, instead of firms bidding for a right to be sole data collector or something similar, piecemeal subsidies for collecting general variables can generate yet more data that works well over time and space, and something short of qualified standards can continue to be left in the hands of innovators.

Section II begins by describing the technical limitations of standardization benefits drawing on examples from natural language processing and agricultural science. Section III turns to legal strategies for encouraging coordinated data collection in the presence of social limitations, and in particular, the role that law and public policy plays in driving down costs among competing groups. Section IV concludes.

II. THE LIMITS OF STANDARDIZATION

It is widely understood that the impressive progress and advances in AI over the past few years have been primarily driven by an exponential increase in computing power, vast production and collection of data, and important breakthroughs in algorithm design.[145] What is less understood is that machine-learning, an important subset of AI,[146] is dependent upon two conditions: (1) that patterns or regularities are observable, and (2) that the environment in which those patterns occur is sufficiently stable.[147] Machine learning loses its advantage when patterns are unseen or the future is uncertain. These conditions, especially the second one, tend to focus on time. Equally important is space.

[145] Anant Maheshwari, et al., *Age of Intelligence*, Microsoft India White Paper, February 2019.

[146] Artificial intelligence is used here as a general term as the ability of machines to improve on their own, after humans set the machine's goals and provide it with some data. Machine learning is used as a subset of AI in which machines look for connections, reach conclusions, or look for more data in ways beyond what its human programmers contemplated. These definitions are not exhaustive, and may not even stand the test of time, but are used here to describe the workflows for constructing predictive models and how those workflows can be streamlined with data collection standards. For a discussion of the optimal division of labor between humans and AI when building predictive models, see Fagan & Levmore, *supra* note 141, §II.

An observable pattern or regularity present in one part of the world may not exist in another. While these two conditions capture this fact, they tend to obscure the importance of environmental consistency across space in order for machine learning to be broadly useful. It is obvious that a predictive model may work in India, but not the United States; the United States can introduce additional and relevant variables that do not exist in India. At the same time, the Indian predictive model may capture variables not present or relevant in the United States, which though critical for accurate prediction in India, offer little predictive power elsewhere. If either the Learnable Regularity Assumption (1) or the Invariance Assumption (2) is violated, then the benefits of standardization become limited. Even wider data sharing across environments is not useful unless it illuminates some aspect of either environment that is stable and measurable.

2.1. DISSIMILAR VARIABLES ACROSS MARKETS

Consider a Natural Language Processing tool developed in India to automate customer service in each of India's twenty-three official languages. For each language, a predictive model might compute several variables, including what the customer says or which questions the customer asks, in order to predict the appropriate output response of the automated call agent. Data collection would surely include customer utterances in the spoken (local) language, and the accuracy of the response given by the predictive model would at least partly depend upon those local utterances. If the model is dependent on language, then its usage is confined to the language of its construction and its market is likely confined to geographic regions where that language achieves critical mass. The model might be extended to account for language-independent features of speech, which could generate predictive capabilities for export, but investment in a language-independent model in this context seems unlikely. One only needs to assume that the economic costs of developing a sufficiently accurate language-independent model in India exceed the costs of developing either type of successful model in the importing locale. This assumption seems reasonable. Language data itself is easy to collect, and its

[147] LESLIE VALIANT, PROBABLY APPROXIMATELY CORRECT 61-62 (2013).

use in developing predictive customer service applications is more straightforward than language-independent features like the time of day when the call takes place, the type of product for which service is required, or the age of the caller.

This does not imply that models which ignore local languages may have important commercial uses that can be profitably exported by frugal innovators. For instance, models that predict caller mood across twenty-three languages based upon a collection of language-independent variables may have important commercial applications in say, the European Union, which itself has twenty-four official languages. The point is that in some cases, language-dependent models for specific customer interactions may get the job done more accurately and at a lower cost, even when developed in a relatively higher cost location. If homegrown models are comparatively efficient, then there is less space for predictive model exporting and fewer benefits accruing from standardization across jurisdictions.

2.2. DISSIMILAR ENVIRONMENTS ACROSS MARKETS

As a second example, consider a model that predicts rice crop yields. This model can be based upon a variety of inputs such as how many seeds and of what type are used in a given amount of space; how much water is absorbed by them; various climatic features such as sunlight, humidity, barometric pressure, and temperature; and so on. It may include features of the soil, such as its density, mineral content, the presence of particular insects and organisms, and the number and type of previous crops grown. A predictive model that incorporates exhaustive features of rice crop yields may include attributes of the farmer such as age, height, and weight, in addition. Many types and combinations of variables can be imagined.

It may appear, on the surface, that if this model were developed in Assam, it may have little value for farmers in Idaho. Assam's growing conditions are different from Idaho's, so what is the need to make comparisons? A robust causal model might direct a farmer to apply lesser water at night than in the morning, no matter what the location is, but even a powerful predictive model may offer little guidance if it has never observed Idahoan features. Only if those features are sufficiently similar to those of Assam, will the model prove useful in Idaho. Equally important is that the *unobserved* features of Assam must be sufficiently similar to their counterparts in Idaho so as to not distort the prediction. If critical features—observed or unobserved—are different, then the Assam model will contain no observations relevant to Idaho to support a prediction there.[148]

Say the model ignores wind velocity, and that Assam experiences higher wind velocities than Idaho. Wind speed is important for rice crops. It increases turbulence in the atmosphere, and as a result, increases the supply of carbon dioxide to plants, thereby accelerating photosynthesis rates. This unmeasured difference between Assam and Idaho, if substantially different, will distort the predictive outcome. But say that all Idahoan rice crops are planted on Idaho's plains. The plains are flat and open and subject to higher wind velocities. Because the model does not measure wind speed, it is only predictively useful to Idahoan rice farmers situated on the plains. The relevant market for the predictive model might be expanded to greater parts of Idaho only if wind speed were measured in Assam. While the measurement of wind speed may increase the cost of developing a predictive model for Assam farmers and offer little economic benefit there, its measurement may be more broadly useful outside of Assam. So long as the additional cost of collecting and testing a model that includes wind speed is worth it, measurement should be undertaken for predictive model exporting. Thus, local features, including geography, can be expected to limit the demand for (and patterns of) standardization of data collection.

III. LAWS ROLE IN DATA COLLECTION

3.1. COORDINATED DATA COLLECTION (PROCESS)

The benefits of coordinated data collection are straightforward. Data portability and

[148] This point is conceptually identical along a time dimension as well. If Assam in 2025 presents sufficiently different patterns or a sufficiently different environment, then a predictive model built in 2019 would be based upon regularities that do not exist anymore.

interoperability reduce the costs of creating predictive models. Coordination reduces (1) duplicative data collection; (2) unnecessary conversion of data formats; (3) translation of communication protocols between routines that collect, organize, and store data; and (4) potentially reduces errors.[149] Law has shown an appetite for standardized information about food, fuel, medicine, appliances, and automobiles—primarily to protect consumers and reduce search costs.[150] Requirements for business-to-business transactions include, among others, the transportation, chemicals, and petroleum products industries.[151] Many of these standards impose requirements on content collection and reporting. In other words, they regulate *what* must be collected and reported. By contrast, process-based standards for coordinating data collection would involve *how* data is collected, organized, and stored. Inasmuch as collection, organization, and storage requires the use of metadata or other variables for portability and interoperability, there will be overlap. Nonetheless, the focus here is on process (and not content) standards. Economists have examined the relationship between standardization and both productivity growth and overall economic growth.[152] In short, standards tend to facilitate competition within standardized markets, which reduce costs and increase product quality, choice, and innovation. On the other hand, standards can lead to long term depression of innovation by reducing choice, increasing market considerations and locking-in an inferior standard.[153]

Thus, while data collection standards may be preconditions for beneficial cross-firm or cross-industry data exchange, their use can lead to social loss. If implemented too soon or too late, opportunities for net increases in growth and innovation may be missed.[154] Even if properly timed, data collection standards may raise barriers for new entrants, stifle

[149] *See* Michal S. Gal & Daniel L. Rubinfeld, *Data Standardization*, 94 NYU L. REV. *12-13 (forthcoming 2019) (discussing the benefits of standardization and noting that standardization can reduce metadata uncertainty).

[150] *See* Saul Levmore & Frank Fagan, *The End of Bargaining in the Digital Age*, 93 CORNELL L. REV. 1469, 1471-72 (2019) (discussing various truth-in-labeling requirements and asserting that law should sometimes require firms to disclose prices to consumers for the same reasons).

[151] *See, e.g.*, NATIONAL INSTITUTE OF WEIGHTS AND MEASURES, *Office of Weights and Measures Programs*, https://www.nist.gov/pml/weights-and-measures/programs (last visited Apr. 2, 2019).

[152] *See, e.g.*, Knut Blind & Andre Jungmittag, *The Impact of Patents and Standards on Macroeconomic Growth: A Panel Approach Covering Four Countries and 12 Sectors*, 29 J. PROD. ANAL. 51, 51 (2008); Joseph Farrell & Garth Saloner, *Standardization, Compatibility, and Innovation*, 16 RAND J. ECON 70, 70 (1985).

[153] Gal & Rubinfeld, *supra* note 149 at *15.

[154] *See generally* FRANK FAGAN, LAW AND THE LIMITS OF GOVERNMENT: TEMPORARY VS. PERMANENT

competition and innovation, and depress the development of predictive models. From this perspective, the imposition or announcement of standards raises challenging policy questions.

If the future is sufficiently certain and the private costs of compliance with data collection standards are low, then endorsement may be worthwhile. On the other hand, if firms benefit from coordination, then a data collection standard might be expected to emerge in the first place as its optimal timing approaches, and its imposition would be unnecessary. If anything, law might announce a standard to encourage coordination.[155] Standards may fail to emerge, however, if incumbents benefit from fragmentation, or collective action problems prevail—including limited knowledge about aggregated data's potential uses, its expected level of integration, or the propensity of others to follow suit.[156] Additional obstacles have been raised in other work,[157] but the main point is that efficient standards may fail to emerge, even with law's endorsement. Of course, the danger of endorsement is that the standard itself is inefficient. Competition among jurisdictions—in particular, a national desire to win the global AI race—may be expected to bring about efficient results, but only if big data is big enough within jurisdictions. Otherwise, federations should be expected to emerge loosely patterned around traditional collective action behaviour, including the concentration of participants as a reflection of organizational and other transaction costs.

3.2. COORDINATED DATA COLLECTION (SUBSTANCE)

Imagine that a genetic variation, which alters the outcome of medicinal treatment, is widespread throughout a national healthcare market, but less so in another. In the market where this variation is uncommon, the collection, organization, and storage of binary data about its presence during treatment may have little impact on the accuracy of predicting local treatment outcomes. Diagnostic trials required by the local administration agency will likely conclude that the inclusion of this variable in testing is of little value. Firms

LEGISLATION (2013).

[155] *See* Richard A. McAdams, *A Focal Point Theory of Expressive Law*, 86 VA. L. REV. 1649, 1649 (2000).

[156] Gal & Rubinfeld, *supra* note 149 at * 23.

may wish to include it anyway, even if its presence is scarce, especially if they anticipate that foreign agencies may require its inclusion in future trials.[158] In this case, the firm might choose to include it in order to facilitate expansion into other markets as a result of its profit maximization calculus.

In this case, the predictive model would then be based upon a greater number of observations and (potentially) more robust to other environments. While process-based coordination of data collection increases the number of observations by essentially reducing aggregation costs, substantive coordination increases the number of observations by providing direct benefits to additional data collection effort. Here, the benefits are clear since the firm increases its capacity for trial testing in other markets. But governments can bring about those benefits through tax incentives or conditional infrastructure funding. Suppose a national government invites security firms to bid on a border checkpoint scanning system, for a specific corridor, that draws heavily on machine learning and data collection. Even if the features and aspects of facial expressions don't provide any additional predictive powers for that particular corridor, the national government will impose broad data collection requirements on bidders in order to build stocks of data for use in other locations or other applications.

IV. CONCLUSION

The benefits of standardization are limited by unobservable patterns and variations over time and space. Even if these technical limitations are few, standardization faces social limitations. When the benefits of data independence are high, creators of predictive models will resist imposed coordination. Even if coordination is efficient, concentrated beneficiaries of the status quo will successfully resist the imposition of standards that weaken their positions. And even if the coordination is efficient for all participants, other collective action problems based upon incomplete information may prevent socially beneficial changes. When these obstacles are surmountable, lawmakers should consider

[157] *See id.*

[158] One can assume that the present inclusion can be controlled in local trials and is useful for later testing.

whether process-based standardization or substantive standardization is efficient. The endorsement of process-based data collection standards entails a social cost, when the standard itself is inefficient, which may be difficult for lawmakers to predict over time. For this reason, an incremental approach realized, for example through substantive data collection requirements of government-funded or tax-incentivized projects, developed and articulated on a project-by-project basis, may minimize errors when the benefits of standards are uncertain. Over time, insofar as benefits become certain and clear, competition among jurisdictions—in particular, the desire of a nation or group of nations to win the global AI race—may be expected to bring about efficient standards, but only if big data is big enough within jurisdictions or across national partners.

It is nonetheless a truism that the benefits of standardization must be discounted, in an expected value sense, by an uncertain future. Standards may or may not generate economies of scale in a given socio-economic environment, but they can be expected to generate centralization and lock-in, while simultaneously inhibiting competition and innovation. This is the danger of centralized standards, either imposed or announced. While economies of scale can lower costs, drive innovation, and enhance welfare generally, all of these benefits depend upon the perfection of the standard over time and space. As a result, standardization arguments can easily be made but difficult to believe after further scrutiny. In other settings, auctions and other price mechanisms serve as scrutinizers, though perhaps here, instead of firms bidding for a right to be sole data collector or something similar, piecemeal subsidies that incentivize the collection of general variables can generate even more data that works well over time and space, and something short of qualified standards, for the form and process of data collection can continue to emerge and remain in the hands of innovators.