When is a Chair not a Chair?
Big Data Algorithms, Disparate Impact, and Considerations of Modular Programming♦

James A. Sherer*

Abstract

Data algorithms associated with human behavior are integral to a variety of business endeavors. Algorithm quality is in turn intimately related to data quality and size, and as collection measures and relative data sizes increase in volume, algorithms commensurately increase in complexity. This makes insight into the inner workings of algorithm function more difficult, which is challenging at a time where both the effects of algorithms and the data associated with their development and utilization are undergoing additional scrutiny. This paper examines these trends, and considers both post-algorithm utilization audits and modular programming approaches as appropriate measures to improve algorithm function and regulatory compliance.

Quality Data (Usually) Leads to Quality Performance

Efficacious algorithms are developed using high quality data sets. Objective data "quality" has therefore emerged—unsurprisingly—as a core component of modern algorithm development.[1] Taking marketing algorithms as a proxy for human behavior-directed algorithms designed to reveal preference, such algorithms are built with view that "historical consumer behavior" and "an individual's or a group's …past choices" are the best predictors of future actions.[2] True historical data is often considered the most important portion of the equation,[3] and bigger is better, where "very large data sets can improve even the worst machine learning algorithms."[4] While there are detractors from this theory who argue that developers are sometimes too trusting of data to the detriment of algorithm development,[5] the majority of current, conventional wisdom trusts in history and volume as proxies for quality.

Quality Data comes from the Real World

The historical portion of high quality data comes from real world experiences. Online sellers who "have large amounts of data about users' past purchases" subsequently "use this data as input" for their marketing algorithms as a matter of course.[6] These data sets support algorithm development premised on the idea that truly rich data sets that identify an individual as well as certain things *about* that individual that the individual might prefer were private or "forgotten," will provide superior granularity when trying to sell a product or service.

Real world quality sources do not stand alone; they do not live in central files that can be deleted at will, or

---

♦ Submitted on May 7, 2107 as a Position Paper for ICAIL 2017 Workshop on Using Advanced Data Analysis in eDiscovery & Related Disciplines to Identify and Protect Sensitive Information in Large Collections ("DESI VII Workshop").

* James A. Sherer is a Partner in the New York office of Baker Hostetler LLP. The views expressed herein are solely those of the author, should not be attributed to his place of employment, colleagues, or clients, and do not constitute solicitation or the provision of legal advice.

[1] Barna Saha & Divesh Srivastava, *Data quality: The other face of big data*, Data Engineering (ICDE), 2014 IEEE 30th International Conference on IEEE (2014).

[2] Joachim Gudmundsson, Pat Morin & Michiel Smid, *Algorithms for Marketing-Mix Optimization*, Algorithmica, 60(4), 1004-1016 (2011).

[3] Michel Banko & Eric Brill, *Scaling to Very Very Large Corpora for Natural Language Disambiguation*, Proceedings of the 39th annual meeting on association for computational linguistics. Association for Computational Linguistics, (2001).

[4] Auren Hoffman, *Where Should Machines Go To Learn?* SafeGraph Blog (Dec. 12, 2016).

[5] Ari Zoldan, *More Data, More Problems; Is Big Data Always Right?* Wired (May 2013).

[6] *See* Gudmundsson *supra* note 2.

that operate under a singular authority. Individual data is instead generated by many different sources (including the Internet of Things, social media, telematics, and biometric activity) and, in part due to those sources—as well as the freedom to share information the digital medium provides—exists in redundant copies and forms around the world. Those sources may be combined to extract even greater value and lessen associated privacy. In one illustrative example, an online entertainment purveyor offered a prize to improve its recommendation algorithm. A research team matched the rich dataset provided with real-world internet "anonymous" reviews to unmask individuals.[7] The researchers effectively linked datasets to key identifiers that would allow merchants to direct advertisements finely tuned to their recipients.

Good does not equal "Right"

Algorithms are used to personalize recommendations, and "[e]very website, every search engine, every video platform…[tries] to surface you the best content out of everything they offer."[8] For purveyors who are selling something *presently*, this works, and works better than any prior means. But these real-world "quality" data sets are mirrors of their times, permeated with prior problems. Algorithm development is not immune: to the contrary, these issues are reflected in the algorithms that arise from real world data sets. Here accuracy is not ideal, as noted in a 2014 Whitehouse Report examining the combination of "detailed personal profiles held about many consumers" and "automated, algorithm-driven decision-making."[9] That practice, it stated, "could lead—intentionally or inadvertently—to discriminatory outcomes" which the report called "digital redlining,"[10] a continuation of the practice of "redlining" where banks would draw boundaries around certain neighborhoods on maps where they would not loan money, excluding generations of minorities.

Some views supporting real data algorithm development argue that this process eliminates human biases from the development decision-making process, but algorithms ultimately comprise the data they work with.[11] Algorithms "trained on historical data" will, for example, "know" that "poor uneducated people (often racial minorities) have a historical trend of being more likely to succumb to [things like] predatory loan advertisements."[12] But that algorithm's output will have an identified customer base, and it will likely work. And if there is a discriminatory issue recognized on the back end, because such an issue is an "emergent property of the algorithm's use" rather than a design consideration, it can be unusually hard to identify the source of the problem.[13]

In this way, high quality data set-generated algorithms can also, even if inadvertently, give rise to differential privacy concerns[14] or the "mosaic effect," whereby "personally identifiable information can be derived or inferred from datasets that do not even include personal identifiers, bringing into focus a picture of who an individual is and what he or she likes."[15] This is an active problem, encountered during the algorithm contest described above. Ultimately, an in-the-closet lesbian mother sued the contest host for invasion of her privacy, claiming that the provider had "outed" her when it disclosed insufficiently anonymous information about her viewing habits (among nearly 500,000 customers) to the public.[16]

---

[7] Ryan Singel, *Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims*, Wired (Dec. 17, 2009).
[8] Jerry Daykin, *The truth about social media algorithms – and why marketers should welcome rather than fear them*, The Drum (April 28, 2016).
[9] Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (May 1, 2014) at 53.
[10] *Id.*
[11] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 California Law Review 671 (2016).
[12] Jeremy Kun, *What does it mean for an algorithm to be fair?* Math Programming (July 13, 2015).
[13] *See* Barocas *supra* note 11.
[14] James A. Sherer, Jenny Le & Amie Taal, *Big Data Discovery, Privacy, and the Application of Differential Privacy Mechanisms*, The Computer & Internet Lawyer 32 (7) (2015).
[15] *See* Executive Office *supra* note 9 at 8.
[16] *See* Singel *supra* note 7.

This is also a passive problem, where individuals are identified as quality targets as the aim of the programs despite the individuals' preferences (who do not want to be "sold" to—or "sold" period). Finally, this concern is also recursive, as once an individual is identified—even if not by name—she is then further incorporated into the set as a person on whom a technique worked. Likewise, marketing algorithms attempt to "understand the contextual drivers" of customer response;[17] the same might hold true, for example, for privacy considerations associated with an individual visiting the same block as a drug treatment facility on a consistent basis.

A Simple Outcome Masks a Complex Process

At heart, an "algorithm is a defined, repeatable process and outcome based on data, processes, and assumptions,"[18] or, stated more simply, one or a combination of "sets of 'if-then' rules."[19] But due to the nature of the data sets upon which complex algorithms are built, and the methods currently utilized to address big data, some of the algorithm approaches are built up in a myriad of interconnected steps through iterations of data analysis, rendering them opaque in nature and incomprehensible, even by those individuals tasked with developing or utilizing them.

Algorithms generated from high-quality data sets are complex and multi-dimensional, incorporate many different analyses concurrently, and are difficult to disambiguate if there are problems like disparate impact. Current architectures can also lead to "opaque encodings" of both programs and their outputs, making it even more "difficult to effectively audit or process" those results.[20] Further, computing environments are imperfect; ambiguity has long been a part of outputs since parallelism introduced the possibility of nondeterminate behavior—the process whereby a software component produces different results, even with the same input, when there is a slight difference in internal events.[21]

These opaque and imperfect algorithmic approaches have led to disparate impact issues, even with attendant human oversight. And individual preferences are increasingly filtered through interactions with Artificial Intelligence introduced at the front end of marketing practices that acts as a gatekeeper for further human interaction. This adds additional distance between what "should" happen to what does, and reinforces current practices in service of immediate results by generating bespoke or individualized social media feeds and related interaction.[22] This algorithmic approach is not just doing it—it is doing it all.

Prior approaches were cognizable, assailable, and could be questioned, but that position has changed. Present-day scientists and mathematicians have taken note, indicating that the design was structured and reviewable as "true," underpinned by "very clear assumptions" and "very logical steps."[23] This approach, at least in theory, provided for instances "where if in fact you were wrong about something, if you were making a step that you thought was logical but it had a flaw, and someone told you that you were making a mistake, you're apt to thank them. You're apt to say, oh, thanks for explaining my mistake, it's saving me time."[24] When addressing why these complex algorithms operate in a certain way, the response may be that the operation represents the way things (likely) are—or that no clear response is available.

---

[17] Utpal M. Dholakia, *The Perils of Algorithm-Based Marketing*, Harvard Business Review (Jun. 17, 2015).
[18] Christopher S. Penn, *Marketers: Master Algorithms Before Diving into Machine Learning* (Feb. 1, 2017).
[19] *See* Dholakia *supra* note 17.
[20] Bill Eidson, Jonathan Maron, Greg Pavlik & Rajesh Raheja, *SOA and the Future of Application Development*, Proceedings of the First International Workshop on Design of Service-Oriented Applications (WDSOA05) (2005).
[21] Jack B. Dennis, *A Parallel Program Execution Model Supporting Modular Software Construction*, Massively Parallel Programming Models, 1997. Proceedings. Third Working Conference on. IEEE (1997).
[22] Ben Rossi, *Marketing faces death by algorithm unless it finds a new code*, The Telegraph (Nov. 10, 2016).
[23] HBR Idea Cast, *When Not to Trust the Algorithm*, Harvard Business Review (October 6, 2016).
[24] *Id.*

Auditing Addresses the Solution but Not the Problem

While a response from the algorithm for its operation might be unavailing, individuals utilizing algorithms to immediate benefit have recognized a need for solutions addressing problematic outputs. One current approach to defending algorithms tests them according to issues that should not arise, auditing results and making inferences based on the data the algorithm uses.[25] The market should support this as well: as noted in one article, "once business leaders realize[d] that they're really putting themselves at risk for using discriminatory hiring practices via an algorithm," they will "actually double-check that this algorithm is legal" which in turn would lead to "a market for that" process.[26]

In this fashion, disparate impact is addressed by auditing results for problematic outcomes. Programmers then work backwards, tweaking foundational data sets until acceptable outcomes are generated. Practitioners utilize this approach for purposes other than disparate impact; they also utilize it to test back-end data end to determine if the algorithms are discovering or unearthing private data that should have been eliminated based on the recommendations from those algorithms.[27]

But post-issue audits seem to be ill-equipped to handle likely regulatory challenges to practices that ignore personal privacy for commercial or other reasons. Specifically, the GDPR 2016/679[28] will replace DPD 95/46/EC[29] within the European Union on May 25, 2018. Among its provisions, GDPR addresses "any information concerning an identified or identifiable natural person,"[30] including "online identifiers" associated with "devices, applications, tools and protocols…or other identifiers."[31] It limits that information collection, review, and use[32] to incorporate individuals' "right to be forgotten."[33] GDPR includes the right to be forgotten, among other restrictions, and presupposes a way in which that right is exercised.

Breaking Up is Hard to Do

Audits seem an imperfect solution to issues involving disparate impact; the incorporation, construction, and unearthing, of personal data; and related concerns. And given the limits of present-day data scrubbing techniques[34] as well as a focus specifically on personal identifiers (such as government identification numbers or the like), automated record sanitization practices also seem to fall short in the face of differential privacy analysis, as noted in a previous article.[35]

The GDPR seems to consider an alternative approach for dealing with algorithmic discrimination: that of algorithm transparency.[36] In *Article 12: Transparent information, communications and modalities for the exercise of the rights of the data subject*, the GDPR further specifies that such information must be provided "in a concise, transparent, intelligible and easily accessible form, using clear and plain language."[37] While this is a

[25] Michael Feldman, Sorrelle A. Friedler, John Moeller, Carlos Scheidegger & Suresh Venkatasubramanian, *Certifying and removing disparate impact*, BIGDATA program (Jul. 16, 2015).
[26] *See* HBR Idea Cast *supra* note 23.
[27] *See* Dholakia *supra* note 17.
[28] Regulation (EU) 2016/679 of the European Parliament – the EU General Data Protection Regulation (GDPR).
[29] Regulation (EU) 1995/42 of the European Parliament – the EU Data Protection Directive (DPD).
[30] *See* GDPR *supra* note 28 at ¶26.
[31] *Id.* at ¶30.
[32] *Id.* at ¶39.
[33] *Id.* at ¶65.
[34] The National Academies of Sciences, Engineering & Medicine, *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions*, Washington, DC: The National Academies Press (2017).
[35] *See* Sherer *supra* note 14.
[36] Bryce W. Goodman, *A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection*, 29th Conference on Neural Information Processing Systems (NIPS 2016).
[37] *Id.*

requirement for an *explanation*, it also seems to require front-end design decisions in order for complex algorithms to comply.

This introduces the concept of advanced algorithm by design rather than effect. This is not a new concept: software and process have an approach to programming and debugging called, at various times, modular programming,[38] functional computing,[39] or functional programming.[40] This process considers well-structured software to be good software, and that structure is paramount as software increases in complexity. Modular computing requires complex systems to be segregated into modular components for evaluative and functional purposes. In short, the approach "avoids the superimposition of a complex syntactic and semantic structure over the simple structure of the basic language"[41] to make "code readable and easy to understand" and accessible in the face of needs to "integrate new team members, fix bugs, and refactor existing code."[42] This paper submits and briefly discusses below how this approach may help algorithm design improve in auditability as well as function.

When a Chair is not a Chair (and a Conclusion)

> *A chair can be made quite easily by making the parts - seat, legs, back etc. - and sticking them together in the right way. But this depends on an ability to make joints and wood glue. Lacking that ability, the only way to make a chair is to carve it in one piece out of a solid block of wood, a much harder task.*[43]

While an algorithm metaphorically carved from a single piece of wood may be effective, its rigidity is not ideal from a practical standpoint, and its opacity prevents its defense. This paper instead asserts that the complex algorithms that birth the concerns discussed herein could benefit from the front-end application of modular programming principles. If done correctly, it was noted, "good software systems are easy to separate into different modules, with the interface between modules being kept relatively sparse and simple."[44] Further, "a *correct* program is one that does exactly what its designers and users intend it to do."[45] This is *may*, rather than *can*; *should* rather than *status quo*.

This is algorithm by design, which considers privacy and disparate impact issues (among others) before setting the algorithm loose to do what it is otherwise intended to do. At a most basic level, algorithms make decisions, and are required to automate processes that are no longer "do-able" by humans at an effective price point. But designed algorithms judge effects other than efficacy and are internally auditable, accompanied by "robust documentation and explanations" for both the algorithm and the purposes for which the algorithm is used.[46]

Certain portions of this type of approach are already underway, including professionals associated with the

---

[38] T.O. Barnett, *Guide for Submission of Papers to the National Symposium of Modular Programming*, Information & Systems Institute, Inc. (May 3, 1968).
[39] Jin Li, Sanjeev Mehrotra & Weirong Zhu, *Prajna: Cloud Service and Interactive Big Data Analytics*, F# - Parallel and Distributed programming (2013).
[40] John Hughes, *Why Functional Programming Matters*, Institutionen för Datavetenskap (1984).
[41] Antonio Brogi, Paolo Mancarella, Dino Pedreschi & Franco Turini, *Modular Logic Programming*, ACM Transactions on Programming Languages and Systems (Jul. 1994).
[42] William Li, Pablo Azar, David Larochelle, Phil Hill, & Andrew W. Lo, *Law Is Code: A Software Engineering Approach to Analyzing the United States Code*, 10 J. Bus. & Tech. L. 297, 308 (2015).
[43] *See* Hughes *supra* note 40 at 3.
[44] *See* Li et al. *supra* note 42, *citing* Steve McConnell, *Code Complete: A Practical Handbook for Software Construction* (2d ed. 2004) at 38.
[45] Zhenjiang Hu, John Hughes & Meng Wang, *How functional programming mattered*, National Science Review 2.3 (2015) at 349-370 (emphasis original).
[46] *See* Penn *supra* note 18.

Fairness, Accountability, and Transparency in Machine Learning (FAT ML) 2015 proceedings.[47] However, strategic approaches focused only on specific process components may miss other less ominous but still impactful ideas, such as the concern that big data technology may "assign people to ideologically or culturally segregated enclaves known as 'filter bubbles' that effectively prevent them from encountering information that challenges their biases or assumptions."[48]

In contrast, utilizing a modular programming strategic approach at algorithm inception may address two issues more broadly: First, algorithms designed in this fashion would be more easily understood, both for how they operate as well as how they can be modified or "fixed." Secondly, a modular programming approach requires additional thought, direction, and foresight at the beginning of the process—it requires a strategy. Strategic considerations involved on the front end, rather than a debugging process on the back, can consider the "should." And the hope is that this will not hold back development, but rather channel it into a more responsible place.

Finally, note that regardless of whatever technique used to combat disparate impact, digitally "forget" someone, or even incorporate modular programming techniques into algorithm design, the approach will undoubtedly itself require the use of computers and advanced algorithmic techniques[49] (incorporating machine learning),[50] due to the big data size, complexity, and nature of the task. As to how that recursive process ultimately plays out, this paper does not present a hypothesis.

[47] FAT ML 2015, *Resources, Relevant Projects, Relevant Events, and Relevant Scholarship* (Jan. 8, 2016).
[48] *See* Executive Office *supra* note 9 at 53, *citing* Cynthia Dwork & Deirdre Mulligan, *It's Not Privacy, and It's Not Fair*, 66 Stan. L. Rev. Online 35 (2013).
[49] *See* Feldman *supra* note 25.
[50] Karthik Guruswamy, *Data Science – Data Cleansing & Curation*, Teradata - Aster Community (Jul. 15, 2016).