Title
Using a narrative-based approach as a safeguard against bias related harm in algorithmic tools and services

Authors
Walter, Paul
Fast, Victoria

Publication Date
2023-09-05

DOI
10.25436/E2C308

Peer reviewed

Using a narrative-based approach as a safeguard against bias related harm in algorithmic tools and services

Paul Walter[1[0009−0005−7465−7325]] and Victoria Fast[1[0000−0002−7093−3864]]

University of Calgary, Calgary, Alberta, Canada
paul.walter@ucalgary.ca

Abstract. The ubiquity of algorithmic tools and services (ATS) in spatial data science has led to increased concerns about the biases they carry. This vision paper explores the biases inherent in ATS, encompassing computational, statistical, human, and systemic biases, and those compounded by multinational corporations. It underscores the imperative to address these biases, advocating a narrative-based approach to counteract them and promote equitable outcomes. This approach not only heightens awareness of embedded biases but also charts a course toward their mitigation.

Keywords: Bias · Deep Learning · Disability Inclusion · Intersectionality · Machine Learning · Mental Model · User Experience

DOI: https://doi.org/10.25436/E2C308

1 Introduction

> "Prejudice is a burden that confuses the past, threatens the future and renders the present inaccessible." [1]

Algorithmic tools and services (ATS), produced by the confluence of big data, machine learning, and geo-AI, are being rapidly integrated into real-world systems, gaining agency, and becoming new actors that can drive cars [12] and predict crime [30, 37]. However, ATS, like all tools, are embodied extensions of people's mental models of how the world works for them [17, 23]. Even spatial technology, though often mistaken for being a physical or digital thing, are mental models encoded into forms. The parts we see or touch are artifacts; physical or digital accompaniments that facilitate the work of the model and help accomplish the aims of the people for whom it was originally intended [23].

The truism that all models are wrong but some are useful, also applies here [7]. As a simplification of reality, part of the model will accurately reflect it, while part of the model, the bias, will not [8]. The resulting tool, whether physical, digital, political, or otherwise, will not only carry the embodiment of the form needed to support the mental model, but it will also carry its bias—as is the case with right-handed scissors for left-handed people. For ATS, this is especially alarming because their unprecedented speed and scalability [30] will amplify bias-related consequences [33].

Amid the growing adoption of ATS in spatial data science, this discussion gains relevance because our field inherits the data-focused [22] "classical" approach prevalent in engineering and scientific disciplines that perceives the world as underlying forms [28], and works to distill them down into immutable, single truths. As geographers, we follow in this tradition by using statistical, vector, and raster models to simplify these forms into orderly, mappable, and knowledge-rich ontologies [6]. Though a useful approach for natural phenomena, a classical approach can overlook a multiplicity of human experience [26, 35], and devalue other ways of knowing tending towards holistic, intuitive, social, and emotional awareness: keys for understanding the people around us. In a geographical context, a classical approach can fail to recognize the existence of other perspectives that contain key entities at finer resolutions [27]. It can also misunderstand at which scale its geospatial processes are operating and therefore aggregate data arbitrarily and inaccurately [25].

A classical approach, when applied to people, can lean too far towards toxic perfectionism, adopting unattainable "normal" bell-curve ideals that can dehumanize [14, 26]. People have intersectional aspects to their identity (such as ability, age, class, culture, education, gender, language, race, sex) which, through their lived experiences, can span a multiplicity of semantic understandings that a purely classical approach could fail to identify because data-first ontologies generally assume one perspective, and therefore limit space for semantic multiplicity. Even seemingly simple concepts can have opposing semantic meanings, as is the case in the field of micro-mobility where "stairs" are understood as both a facilitator for walking pedestrians, and a barrier to rolling ones.

As ATS becomes increasingly woven into our research and computational environments, complexity grows in tandem, necessitating the creation of nuanced, human-centric, approaches that recognize outliers, confront bias, and cultivate an inclusive framework for spatial data science. Leveraging insights from the fields of user experience design [24], business service design [20], and software development [11, 19]—all of which emphasize a diversity of human-centered perspectives—this vision paper serves as a catalyst for reflection and action. Starting with an exploration of biases in ATS, including computational, statistical, human, and systemic biases (Section 2.0), the paper then delves into strategies to overcome these challenges (Section 3.0), culminating in a conclusion that charts a path towards a future where ATS is adopted on more equitable terms (Section 4.0).

2 Biases in ATS

The biases in ATS are like layers in an iceberg. Each layer weighs down the previous ones, compounding the difficulty of its detection, and increasing the reach of its consequences [33].

2.1 Computational and Statistical Biases

Computational and statistical biases (CSBs), the top layer and the easiest of the three types of biases to detect, are currently the only layer being addressed in ATS [33].

Because of ATS reliance on statistical methods, they often disregard "noise" [34], which results in three CSB bias subtypes: data bias, sampling bias, and algorithmic bias [39]. This proves hazardous for equity-deserving populations because their heterogeneity of lived experience is expressed in terms of statistical "outliers," a documented weakness in ATS [14].

Instances of CSB bias are evident in the inability of self-driving cars to recognize wheelchair users as pedestrians due to physical form outliers, resulting in serious safety concerns [38]. Additionally, ATS support frameworks often utilize culturally dominant languages in their usage and encodings, frequently disregarding non-dominant letters. This oversight may lead to the exclusion of large portions of the population and undermine the efficacy of these tools [32].

2.2 Human Bias

Human bias (HB), just under the waterline, is harder to detect [33]. It encompasses the mental shortcuts a person uses to deal with the complexity of reality [8]. Examples of biases can include confirmation bias, where individuals tend to seek information that aligns with their preexisting beliefs or hypotheses [33], as well as racial bias [3].

Significantly exacerbated by a lack of understanding of digital agents, human bias is particularly dangerous due to the innocuous nature of ATS. These tools subtly exert influence over important decisions within seemingly commonplace systems [36, 37]. The agents operating within their mental constructs often remain opaque [31, 37], inadvertently perpetuating harm against individuals whose demographic attributes diverge from those of ATS creators.

These human biases are not more recognized because, in part, our languages have polysemes; contextually specific words that change semantic meaning depending on their situation [9], like how "disability" is understood differently in medicine, law, politics, and people's lived experience [13]. Polysemes are linguistic points of contact for these overlapping narratives. This overlap can lead to narrative occlusion for cultural minorities. Narratives are (linguistic) tools, and like all other human tools—legislative, political, artistic, technological, and civic (the built environment)—are merely encoded mental models. This scenario is particularly concerning for individuals who lack digital literacy, hampering their ability to identify and report these harms due to the digital divide [21].

2.3 Systemic Bias

The deepest layer of the iceberg is systemic bias, the hardest and most complicated to address [33]. It is the same as human bias except aggregated over larger extents: geographic, temporal, and by the number of actors in a system (i.e., institutions). Its invisibility contributes to its pervasiveness and power. This relates to how ATS are placed in society in systems that have long since embodied misogyny, racism, and ableism [10].

The systemic bias is compounded through the influence of large multinational corporations (MNCs), the few organizations that have the resources to create large ATS. MNCs often adopt short-term, risky, and value-seeking strategies to gain market hegemony in a global "do or die" market economy. Furthermore, MNCs may not respect or be subject to a country's sovereign laws [29]. This means the rights of individuals or groups can be compromised without fear of reprisal. The ATS they make can, and have, led to unfortunate and unintended consequences, such as with Facebook pursuing short term gain over long term political stability as evidenced in the Cambridge Analytica scandal [16].

3 How do we overcome these biases and risks?

As geographers and spatial data scientists, our role in addressing bias risks begins with recognizing that ATS are not neutral forces, and instead are shaped by the language, values, and beliefs of the cultural, political, and economic systems that underpin their creation [5, 35]. Our first step is to acknowledge these biases and their implications before we embark on strategies to mitigate them. Stop and ask: who runs, controls, and trains the proprietary training data and libraries we rely on?

Drawing insights from human factors research and human-computer interaction research, one solution is adopting a bottom-up narrative-based approach, a technique software engineers have been employing for over three decades [19]. This approach starts with the collection of narratives from domain experts explaining their aims. From these narratives, we can identify the interactions that need to occur. The interactions then are decomposed into actors and their associated tasks—much like a script in a play—to form use cases. The use cases get verified by the domain experts, then forward engineered into software. By contextualizing and validating the user experience in this bottom-up way, we can more deeply understand the in-situ experiences of equity-deserving populations and address their needs.

An additionally important aspect in this approach is that the actors can be generalized into personas which can act as benchmarks for ATS creation for both upstream (pre-ATS development) and downstream (post-ATS development), ensuring sensitivity to outliers and marginalized groups. This approach aligns with the recommendations of Tuan [35] and Rundstrom [32], emphasizing the importance of inclusive data collection that values the perspectives of significant "outliers" in decision-making processes.

More tangibly, open-source toolkits like IBM's "AI Fairness 360 toolkit" [4] can help identify and mitigate statistical bias within ATS. Simultaneously, we advocate using established narrative based methodologies like Object-Oriented Analysis and Design (OOAD) methodologies and Unified Modeling Language (UML) [18] to meticulously analyze the architecture of both ATS datasets and outputs. It is notable that these methodologies are already in use by ATS producers, offering a shared communication channel for collective bias reduction efforts.

Expanding upon the narrative-based paradigm, we propose active engagement with domain populations during the development phase. This approach amplifies the voices of those most impacted by biases. Two strategies emerge: first, by assigning value to "outliers" through standardized personas and their documented needs; second, by ensuring that ATS are sensitive to outliers in both data collection and output. By collectively embracing these strategies, we move closer to realizing the vision of fairer and more equitable ATS development and adoption.

4 Conclusion: Moving into the future on more equitable terms

In response to Geoffrey Hinton's departure from Google, an op-ed in the Chicago Tribune highlights the need for real guidance on AI, instead of vague alarmism [15]. We propose that real guidance can be found in empowering a broader spectrum of individuals to engage in discussions about the biases and risks inherent in algorithmic tools and services. Our approach underscores the understanding that ATS are intricately encoded narratives, firmly entrenched within specific ontological frameworks and tailored for distinct mental models, inevitably carrying biases, and therefore it matters whose mental models are considered. Participation is key, and tokenistic positions regarding ATS are unacceptable [2].

To effectively counter biases introduced in ATS, we must squarely address the semantic foundations underpinning these systems. This entails adopting balanced approaches that can handle the diverse ontologies arising from overlapping user experiences. As researchers actively involved in the development and adoption of ATS, it remains paramount to remain conscious of the organizations we endorse, the problem-solving approaches and tools we choose, and the inadvertent biases we introduce. Deliberately recognizing, evaluating, addressing, and overcoming these biases within our evolving ATS-centered research initiatives is pivotal.

References

1. Angelou, M.: All God's Children Need Traveling Shoes. New York: Random House (1997)
2. Arnstein, S.R.: A Ladder of Citizen Participation. Journal of the American Institute of Planners **35**(4), 216–224 (Jul 1969). https://doi.org/10.1080/01944366908977225
3. Banaji, M.R., Greenwald, A.G.: Blindspot: Hidden Biases of Good People. Random House Publishing Group (Aug 2016)
4. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards,

J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias (Oct 2018)
5. Blue, G., Rosol, M., Fast, V.: Justice as Parity of Participation. Journal of the American Planning Association **85**(3), 363–376 (Jul 2019). https://doi.org/10.1080/01944363.2019.1619476
6. Bowker, G.C., Star, S.L.: Sorting Things Out: Classification and Its Consequences. The MIT Press (Aug 2000). https://doi.org/10.7551/mitpress/6352.001.0001
7. Box, G.E.P.: Science and Statistics. Journal of the American Statistical Association **71**(356), 791–799 (Dec 1976). https://doi.org/10.1080/01621459.1976.10480949
8. Cabrera, D.: Systems Thinking Made Simple: New Hope for Solving Wicked Problems. Odyssean Press (2015)
9. Durrheim, K., Schuld, M., Mafunda, M., Mazibuko, S.: Using Word Embeddings to Investigate Cultural Biases. British Journal of Social Psychology **62**(1), 617–629 (Jan 2023). https://doi.org/10.1111/bjso.12560
10. Elwood, S.: Toward a Fourth Generation Critical GIS: Extraordinary Politics. ACME: An International Journal for Critical Geographies **21**(4), 436–447 (May 2022)
11. Evans, E.: Domain-driven Design: Tackling Complexity in the Heart of Software. Addison-Wesley Professional (2004)
12. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A Survey of Deep Learning Techniques for Autonomous Driving. Journal of Field Robotics **37**(3), 362–386 (Apr 2020). https://doi.org/10.1002/rob.21918
13. Grue, J.: Now You See It, Now You Don't: A Discourse View of Disability and Multidisciplinarity. Alter **11**(3), 168–178 (Jul 2017). https://doi.org/10.1016/j.alter.2017.05.002
14. Hendren, S.: What Can a Body Do? How We Meet the Built World. Riverhead Books, New York, illustrated edition edn. (Aug 2020)
15. Hosseini, M.: We need real guidance about AI, not vague alarmism (May 2023), from the Chicago Tribune
16. Isaak, J., Hanna, M.J.: User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. Computer **51**(8), 56–59 (Aug 2018). https://doi.org/10.1109/MC.2018.3191268
17. Jacobson, I.: Object-oriented software engineering: a use case driven approach. Information and Software Technology **34**(12), 825–826 (Dec 1992). https://doi.org/10.1016/0950-5849(92)90126-A
18. Jacobson, I., Booch, G., Rumbaugh, J.: The Unified Software Development Process: The Complete Guide to. Addison-Wesley Professional, Reading, Mass, 1st edition edn. (Feb 1999)
19. Jacobson, I., Spence, I., Kerr, B.: Use-Case 2.0: The Hub of Software Development. Queue **14**(1), 94–123 (Jan 2016). https://doi.org/10.1145/2898442.2912151
20. Lewrick, M., Link, P., Leifer, L.: The Design Thinking Playbook: Mindful Digital Transformation of Teams, Products, Services, Businesses and Ecosystems. Wiley, Hoboken, 1 edn. (Jun 2018)
21. Mackinnon, D., Burns, R., Fast, V.: Digital (In)justice in the Smart City. University of Toronto Press (2023)
22. Miller, H.J., Goodchild, M.F.: Data-Driven Geography. GeoJournal **80**(4), 449–461 (Aug 2015). https://doi.org/10.1007/s10708-014-9602-6
23. Norman, D.: The Design of Everyday Things: Revised and Expanded Edition. Basic Books (Nov 2013)

24. Norman, D., Stappers, P.J.: DesignX: Complex Sociotechnical Systems. She Ji: The Journal of Design, Economics, and Innovation **1**(2), 83–106 (2015). https://doi.org/10.1016/j.sheji.2016.01.002
25. Openshaw, S.: The Modifiable Areal Unit Problem. No. 48 in Concepts and Techniques in Modern Geography, Geo Books, Norwick, UK (1984)
26. Perez, C.C.: Invisible Women. Abrams Press (Mar 2020)
27. Piantadosi, S., Byar, D.P., Green, S.B.: The Ecological Fallacy. American Journal of Epidemiology **127**(5), 893–904 (May 1988). https://doi.org/10.1093/oxfordjournals.aje.a114892
28. Pirsig, R.M.: Zen and the Art of Motorcycle Maintenance: an Inquiry Into Values. New York, Morrow (1974)
29. Ramasastry, A.: Corporate Complicity: From Nuremberg to Rangoon - An Examination of Forced Labor Cases and Their Impact on the Liability of Multinational Corporations. Articles **20**, 91 (Jan 2002)
30. Richardson, R., Schultz, J., Crawford, K.: Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (Feb 2019)
31. Rudin, C.: Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead (Sep 2019)
32. Rundstrom, R.A.: GIS, Indigenous Peoples, and Epistemological Diversity. Cartography and Geographic Information Systems **22**(1), 45–57 (Jan 1995). https://doi.org/10.1559/152304095782540564
33. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P.: Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. Tech. rep., National Institute of Standards and Technology, Gaithersburg, MD (Mar 2022). https://doi.org/10.6028/NIST.SP.1270
34. Trewin, S.: AI Fairness for People with Disabilities: Point of View (Nov 2018). https://doi.org/10.48550/arXiv.1811.10670
35. Tuan, Y.F.: Cultural Pluralism and Technology. Geographical Review **79**(3), 269 (Jul 1989). https://doi.org/10.2307/215572
36. Varshney, K.R., Alemzadeh, H.: On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products (Aug 2017), `http://arxiv.org/abs/1610.01256`
37. Wexler, R.: Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. SSRN Electronic Journal (2017). https://doi.org/10.2139/ssrn.2920883
38. Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, L., Mills, M., Morris, M.R., Rankin, J., Rogers, E., Salas, M., West, S.M.: Disability, Bias, and AI. Tech. Rep. MSR-TR-2019-38, AI Now Institute (Nov 2019)
39. Xu, J., Xiao, Y., Wang, W.H., Ning, Y., Shenkman, E.A., Bian, J., Wang, F.: Algorithmic fairness in computational medicine. eBioMedicine **84**, 104250 (Oct 2022). https://doi.org/10.1016/j.ebiom.2022.104250