Algorithms in College Admissions

Benjamin Lira[a,1]

[a]University of Pennsylvania

This manuscript was compiled on October 11, 2022

What is the best way to incorporate algorithms in complex selection tasks? At present, college admissions relies on a process known as holistic admissions, whereby admissions officers review the whole of an applicant's file, and come to an integrated "gut feeling" judgment of whether this student should receive admission or not. Research in human judgment suggests that such an approach to decision making is likely riddled by noise. While using algorithms in such settings would help reduce unreliability concerns, research on algorithm aversion suggests that admissions officers, and the public in general might be opposed to such an approach. In this investigation we test whether combining the judgement of humans and algorithms yields increases in the accuracy of judgements, while preserving the acceptability of such an approach.

Algorithms in Judgement and Decision Making | Clinical and Actuarial Prediction | Algorithm Aversion

"Where there is judgment, there is noise—and usually much more of it than you think"[1,2]. Each year, vast numbers of students apply to colleges in the U.S. Compared to other countries, the application process in the U.S. entails vast amounts of information, and particularly qualitative information, such as personal essays, letters of recommendation, and reports about extracurricular activities. Every year, universities hire human readers to qualitatively evaluate these applications in a process called holistic admissions[3]. The process is resource intensive, both in terms of time and expense[4,5]; difficult to audit for bias, given that it is based on qualitative assessments, and—of particular relevance to this investigation—prone to high levels of noise: that is unsystematic variation between raters (i.e., level noise), between similar or identical applicants being evaluated by different admissions officers (i.e. pattern noise), and even when a single officers reviews the same file in different moments (i.e., occasion noise).

Algorithms can eliminate this noise because they provide the same output for any given input. However, algorithm aversion[6] makes the implementation of such approaches difficult. A case study of one of such failures is that of the GRADE system, deployed at UT Austin's Computer Science graduate admissions department[7]. The objectives of this investigation are twofold. First, we conduct the first large scale noise audit of college admissions. Second, we test which of the following 4 approaches results in the lowest levels of error while maintaining appropriate levels of social acceptability.

Holistic review. In this section I would explain what holistic review is[3,8,9] and how it is usually conducted in American universities. I would highlight the most relevant problems such as the absence of rubrics and the lack of mechanical aggregation.

Problems in human judgment. In this section, I would present evidence of noise in professional judgment in areas that are far more systematized than holisitc admissions such as medical diagnoses[10,11] and judicial decisionmaking[12]. In the particular context of college admissions there is an interesting paper by Uri Simonsohn that shows that shows that admissions officers tend to give more weight to academic attributes on cloudy days, and more weight to non-academic attributes in sunny days[13].

Algorithm aversion and human-centered AI. In this section, I will start by explaining the failed attempt to use algorithms in post-graduate admissions in the computer science department at UT Austin[7]. I will explain the backlash against this system based on the theory of algorithm aversion[6], and in particular based on concerns about fairness and bias; which are likely the central concerns surrounding the use of algorithms in college admissions, given recent scandals in hiring[14], predictive policing, and healthcare.

As a counterpoint to these potentially justified fears surrounding the use of algorithms in high-stakes situations, I will present human-centered artificial intelligence (HCAI) an approach to intelligent systems that argues for keeping humans in the loop, maintaining final decisions in the hands of humans, and aiming for algorithms that are fair, transparent, and explainable[15,16].

Clinical and actuarial judgement, and their combinations. In this section I will briefly present the classic debate between clinical vs. actuarial judgment[17]. I will highlight that the status quo of admissions in the US clearly relies on the clinical method, as opposed to methods used in other countries.

I will then highlight past research showing how combining actuarial and clinical judgement usually fare in terms of error reduction[18–21].

Current investigation. This investigation has two objectives. The first is to conduct a noise audit of college admissions. To this end, online participants (Sample 1), and college admissions officers (Sample 2) will review application materials of the 2008/2009 application cycle over a number of days. This will allow us to test the reliability of their judgments across cases, across judges, and across situations. Because this data is merged to ground-truth outcome data on college graduation, we will also be able to test for evidence of bias in their judgments (i.e., do people or admissions officers tend to systematically over or underestimate the likelihood of certain groups to graduate). I hypothesise that even though participants will be more worried about bias, noise will be a larger contributor to error in prediction. That said, there will be large amounts of noise in the ratings (Krippendorf's alphas for interrater reliabilities will be lower than .60). Despite high levels of error, the acceptability of the status quo will be high, with participants suggesting that it is an adequate way to carry out college admissions.

[1]To whom correspondence should be addressed. E-mail: bliraupenn.edu

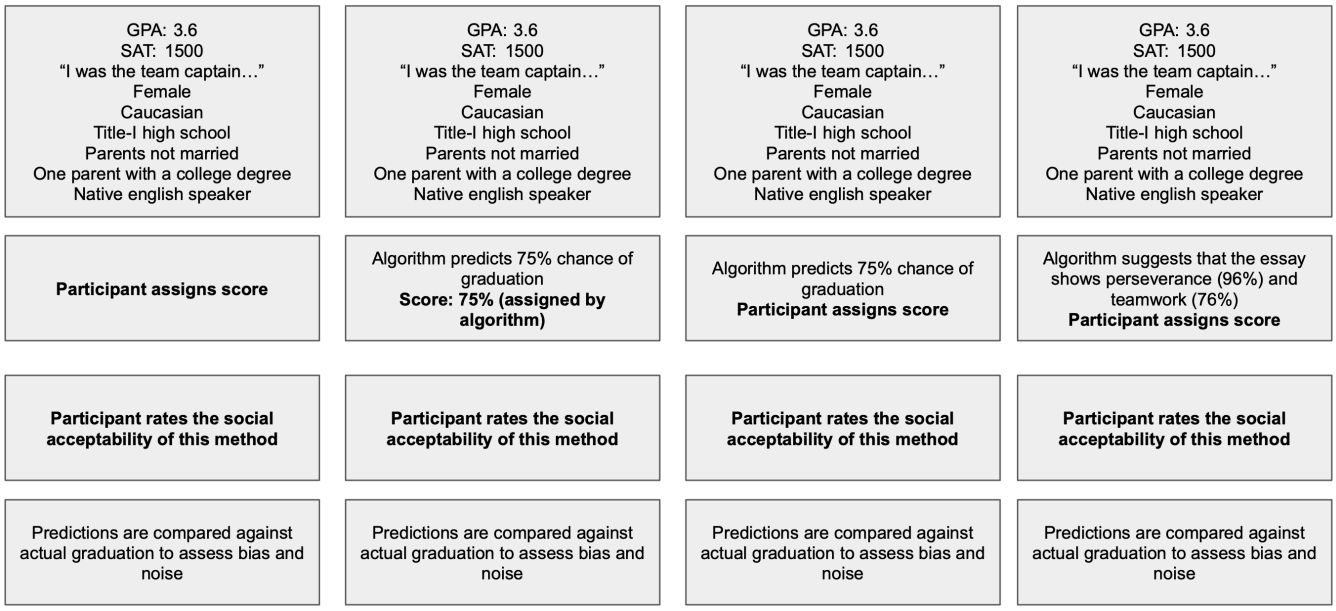

Fig. 1. The four conditions of the study: Holistic review, pure algorithm, human + pure algorithm, bootstrapped mediation

The second objective of the present research is to evaluate how different ways of using algorithms might reduce both bias and noise in college admissions. Aside from the baseline condition described above participants will be assigned to rate the likelihood of graduation of applicants and the social acceptability of: (1) a model that simply predicts the likelihood of graduation based on application materials (just acceptability ratings), (2) seeing application materials and the algorithm's generated predicted probability of graduation, (3) seeing application materials, and algorithms predictions of mediated variables (e.g., a model based on admissions officers suggests that this applicants essays shows that they have high levels of leadership and perseverance). We predict that the more control the human has over the process and the less automatic the judgment feels, the more the method will be deemed socially acceptable, but the less accurate the method will be. In comparing the acceptability and accuracy of these 4 different approaches, we intend to make visible the tradeoffs between accuracy, reliability, fairness, and social acceptability.

Methods

Participants and procedure. Sample 1 will be composed of 1000 prolific participants ($n = 250$ per condition). They will be sampled using the prolific platform, and we will ask them to log into a survey to rate 15 applicants each day over the course of a week. They will see most applicants only once, but a subset of applicants they will see twice to calculate occasion noise (i.e., differences in judgment when the same judge sees the same case twice in different occasions). After rating the applicants, they will reply to a set of items addressing how comfortable they would feel if college admissions were carried out in that manner.

Sample 2 will be comprised of college admissions officers. Given the difficulty of recruiting this population, we will use snowball sampling and simply invite them to review as many applications as they can manage.

Measures. We will provide real application information (i.e., high school GPA, SAT scores, a short essay about extracurricular activities, gender, race/ethnicity, type of high school, parents' marital status, parental education, and English language learner status).

Participants will judge these applications (accompanied by different algorithmic outputs, depending on the condition) by providing a percentage score ("How likely do you think it is that this student will go on to gradate within 4 years if admitted to college?" 0% - 100%).

After completing their ratings, participants will complete a short measure of the social acceptability of the method of admissions they just participated in. The items will be "I think this method is fair", "I think this is a good way for colleges to admit students", "I think this method probably results in colleges admitting students who do go on to graduate", and they will be responded to in a 1 (*completely disagree*) to 7 (*completely agree*) scale.

At the end of their ratings, we will also ask participants to estimate the degree to which they believe their ratings will be reliable. When you saw the same applicant in different days, how far off on average do you think your ratings will be?, When other people see the same applicants as you, how far off do you think their rating will be compared to yours?, If we were to take your average predicted probability of graduation across cases, how different do you think it will be from other participants average predicted probability of graduation across cases?

Then we will calculate the actual level, pattern and occasion noise across conditions. We will calculate statistical bias as the difference between average graduation rates, and average predicted probabilities of graduation, and evaluate fairness by the differences between the actual levels of graduation vs. the predicted levels of graduation for each demographic category. Actual ground truth four-year graduation data is available through the National Student Clearinghouse dataset [22]. The algorithmic scores of character traits are derived from the

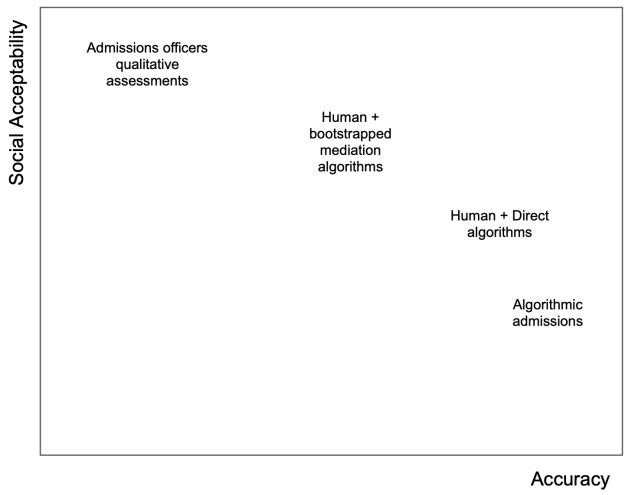

Fig. 2. Tradeoffs between accuracy and social acceptability of different combinations of human and algorithmic judgment of college applications

application of a model that measures personal qualities from these essays [23]

Results

The status quo of holistic admissions has high amounts of noise and bias. We show that there are high levels of noise and bias in our simplified simulation of holistic admission.

Breaking error into noise and bias. How much of error in prediction is caused by bias, and how much of it is caused by noise.

Decomposing noise into level, pattern, and occasion noise. We show the relative contributions of the different kinds of noise to unsystematic error.

Using algorithms significantly reduces both noise and bias. We compare the results of the noise audit for the baseline case, against our three different situations that feature algorithms in different levels of centrality.

Tradeoffs between accuracy and social acceptability of algorithms. We show the relative tradeoffs in terms of accuracy and social acceptability of algorithms. We show that counterintuitively, people and college admissions officers prefer methods that yield higher levels of error. We follow up with participants and capture whether they change their position on the acceptability of these methods after seeing the relative tradeoffs. People adjust their level of acceptability, but not as much as they should.

Discussion

In this investigation we tested how can human judgment be combined with algorithmic recommendations of admission. We show that there are high levels of bias and noise in the status quo of holistic admissions and that noise contributes more than bias to these errors. We show that there are relative differences in the sources of this noise, with pattern noise being the largest contributor to noise. We showed that

Implications. As far as we know, this study would be the first comprehensive noise audit of college admissions. Whereas prior efforts have shown that college application essays encode family income (at least as much as SAT scores) [24], no study to date has evaluated level, pattern, and ocassion noise, together with bias, and fairness in college admissions. We would compare the amount of noise in college admissions with other domains like underwriting, medical diagnoses, and jury decisions here.

A second implication in this study is in the area of human-computer interaction and algorithm aversion. We show how different features of algorithms lead to tradeoffs in acceptability vs. performance.

A practical implication for this research is its obvious applicability to college admissions. With rising numbers of applicants each year, increased social concerns about fairness and access, and understaffed admissions offices, finding ways of increasing efficiency through the use of algorithms without causing public opinion debacles seems like a priority.

Limitations. Several limitations of this study should be noted.

First, in this exercise we made the assumption that in admitting students, colleges are selecting those students who they believe are most likely to graduate. Of course, this is not the case, and colleges have other concerns when selecting students, like building a diverse class, notions of cultural fit, among others. However, graduation is a readily available outcome that is more clearly defined than other aspects colleges aim for. If colleges were indeed maximizing for outcomes other than future graduation, then these methods could be generalized to them.

Second, locating admissions officers is difficult.

Third, our results may be underestimating pattern noise. Because the study was conducted over a week, it is possible that participants remember the scores given to applicants when seeing them for a second time and simply try to match their initial estimates. This concern is partly addressed by not telling participants that they will see some applicants more than once (allowing us to test for a sensitivity analysis using only the first time they saw a repeated applicant). The use of a continuous rather than a binary scale also protects against this concern.

Conclusion. Human judgment is riddled by noise and college admissions is not an exception to this rule. Modern advances in machine learning, and particularly in models that account for textual data (that was difficult to incorporate prior to these advances), afford the possibility of augmenting human judgment and nudging it in the direction of less noise, less bias, and more equity. While algorithmic aversion, and the multidimensional nature of college admissions in the real world make it unlikely that college admissions will be fully automatized any time soon, teams of humans and AI working together might provide the most a

1. D Kahneman, O Sibony, CR Sunstein, *Noise: A Flaw in Human Judgment.* (Harper Collins), (2021).
2. D Kahneman, AM Rosenfield, L Gandhi, T Blaser, Noise. How to Overcome the High, Hidden Cost of Inconsistent Decision Making. *Harv. Bus. Rev.* pp. 38–46 (2016).
3. A Alvero, et al., AI and Holistic Review: Informing Human Reading in College Admissions in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* (ACM, New York NY USA), pp. 200–206 (2020).
4. E Hoover, Colleges Seek 'Noncognitive' Measures of Applicants; Admissions offices want to know about traits, like leadership, initiative, and grit, that the SAT doesn't test. *The Chron. High. Educ.* **59** (2013).
5. M Korn, Some Elite Colleges Review an Application in 8 Minutes (or Less). *Wall Str. J.* (2018).

6. BJ Dietvorst, JP Simmons, C Massey, Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2015).
7. A Waters, R Miikkulainen, GRADE: Machine Learning Support for Graduate Admissions. *Proc. 25th Conf. on Innov. Appl. Artif. Intell.* p. 8 (2013).
8. National Association for College Admission Counseling, Character and the college admission process, Technical report (2020).
9. AL Coleman, JL Keith, Understanding Holistic Review in Higher Education Admissions, (College Board, New York), Technical report (2018).
10. D Jaramillo, Radiologists and Their Noise: Variability in Human Judgment, Fallibility, and Strategies to Improve Accuracy. *Radiology* **302**, 511–512 (2022).
11. CF Mullins, J Coughlan, Noise in medical decision making: a silent epidemic? *Postgrad. Med. J.* pp. postgradmedj–2022–141582 (2022).
12. S Srinivasan, Ph.D. thesis (Harvard University) (2022).
13. U Simonsohn, Clouds make nerds look good: field evidence of the impact of incidental factors on decision making. *J. Behav. Decis. Mak.* p. 10 (2006).
14. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018).
15. MO Riedl, Human-centered artificial intelligence and machine learning. *Hum. Behav. Emerg. Technol.* **1**, 33–36 (2019).
16. B Shneiderman, Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interact.* pp. 109–124 (2020).
17. RM Dawes, D Faust, PE Meehl, Clinical versus actuarial judgment. *Science* **243**, 1668–1674 (1989) Publisher: American Association for the Advancement of Science.
18. J Sawyer, Measurement and prediction, clinical and statistical. *Psychol. bulletin* **66**, 178 (1966) Publisher: American Psychological Association.
19. RC Blattberg, SJ Hoch, Database models and managerial intuition: 50% model+ 50% manager. *Manag. science* **36**, 887–899 (1990) Publisher: INFORMS.
20. A Graefe, JS Armstrong, AG Cuzán, RJ Jones, Combined forecasts of the 2008 election: The Pollyvote. *Foresight* (2009).
21. A Graefe, JS Armstrong, RJ Jones Jr, AG Cuzán, Combining forecasts: An application to elections. *Int. J. Forecast.* **30**, 43–54 (2014) Publisher: Elsevier.
22. SM Dynarski, SW Hemelt, JM Hyman, The missing manual: Using National Student Clearinghouse data to track postsecondary outcomes. *Educ. Eval. Policy Analysis* **37**, 53S– 79S (2015).
23. B Lira, et al., Using Human-Centered Artificial Intelligence to Assess Personal Qualities in College Admissions (2022).
24. A Alvero, et al., Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Sci. Adv.* **7**, eabi9031 (2021).