Andrew Williams & Suhail Bhat

Algorithmic Investigation - Reporting Plan

April 19, 2020

1. A description of your story's main strategy or angle, by description the principal *type of analysis* you intend to apply to your algorithmic system. Will it be a formal analysis (e.g. looking at elements like accuracy, validity or false-positive rates)? Or will it be an audit analysis (e.g. gathering exemplar data and attempting to infer how the system works)? You might also focus on accountability (e.g. who is responsible for mistakes), or appropriateness (in the context of relevant social or legal norms, should an algorithm even be applied in this domain)?

In our DNA Testing Algorithmic Investigation, we will see elements of formal, accountability and appropriateness analysis due to the wide variety of different types of DNA testing. We’ll primarily be focusing on accuracy. Appropriateness will also be a piece of the investigation due to all of the ethical implications currently surrounding DNA testing like FamilyTreeDNA’s betrayal of customers by sharing their data with the FBI and how people are now being denied insurance based on results from DNA testing kits.

There is also a more complex question about the transparency of DNA testing algorithms that are used to generate DNA patterns based on a small sample. The use of algorithmic models is the most recent development in DNA testing that was designed by TrueAllele. Using probability models, its software would generate missing parts of DNA, but it never made public its source code or data.

---------------------------------------------------------------------------------------------------------------------

2. A list of data sets (at least 2) needed for your investigation, and how you would obtain them (e.g. FOIA, API, scraping, user-contributed data collection). Each of these must be things that *exist* (e.g. records you know are collected by the agency you will FOIA, data that the users you will seek out would actually have as a result of their interaction with the system in question). For each, list at least 2 things you would look for in the resulting data, and what approach you would use to identify those potential patterns.

1. **GEDmatch data** | Essentially a large, public, cross-platform DNA database where people willingly input their DNA test results from platforms like 23andMe and AncestryDNA to match with relatives who have taken tests outside of their initial test. We would obtain this data through the user-contributed data collection and scraping.
   1. We want to see what DNA data is being targeted to make matches and which genetic markers are used.
   2. We also want to see how often GEDmatch results are updated. The companies they get the data from being updated constantly and these updates can radically change their initial ancestry results. Do these changes extend to GEDmatch too?
2. **TrueAllele** | We would ask TrueAllele to share some of their data with us. If they do, that would be a milestone. But it would be a challenge to squeeze that out of them. If that doesn’t work, we would submit a FOIA request with the Department of Justice asking for the number of cases in which conviction was based on DNA matches and the cases in which those convictions were overturned. There is anecdotal evidence that shows hundreds of wrongful convictions were made based on DNA matches. The Innocence Project helped overturn more than 100 cases in which DNA matches were evidently used to convict defendants. If the justice department denies our requests, we would ask the Innocence Project to share their data and findings with us.

---------------------------------------------------------------------------------------------------------------------

3. A list of stakeholders(at least 2 for each *stakeholder group)*, and how you would reach them. For private individuals affected by the system, you should describe how you will locate them (e.g. social media, legal cases etc.). For designers and/or owner stakeholders, you should be able to identify them by name. List at least 3 questions you would ask of each.

***Individuals***We’ve had the most luck contacting regular people through social media. My preference would be to start with Facebook and then Instagram. Maybe then we can look at some Amazon reviews and then see how

1. A FamilyTreeDNA user (would settle for another testing company user, but FamilyTreeDNA is preferable)
   1. Why did you decide to take a DNA test?
   2. Before submitting your DNA, did you hesitate at all? Why?
   3. Did you look at the company’s privacy policy?
   4. How do you feel about FamilyTreeDNA giving your DNA to the FBI?
   5. What do you think about the “if you have nothing to hide” argument?
2. Someone who has taken a DNA test only for them to be denied insurance coverage later or offered insurance with higher rates.
   1. When did you first take the DNA test?
   2. What was the reason?
   3. Was this a medical DNA test through a doctor or an at-home consumer test?
   4. Did you ever think it would impact your ability to be insured?
   5. How has this affected your family’s ability to get covered since they share your DNA?

**Designers/Owners**

We have a few interviews lined up with people who run DNA testing labs. We would reach out to Mark Perlin via his email and phone. In the past, he has spoken to reporters. We would also reach out to the FBI's press office and ask them to put us in touch with experts in their DNA Casework department.

1. Mark Perlin, the CEO of the DNA-testing firm Cybergenetics and designer of TrueAllele
   1. How does the model work?
   2. What kind of markers does it look for?
   3. What is the accuracy of your probability model?
   4. What kind of sample data was used in the design and how did you validate it?
   5. Why do you think an algorithmic system will make a better decision than a scientist?
2. Sean Crabbe, an Intelligence Analyst at FBI who has worked as a DNA analyst, would contact him through LinkedIn or press office.
   1. How has the role of DNA evidence evolved in the justice system?
   2. Can a DNA test implicate the wrong person? If it does, what are the different ways it can occur?
   3. Is human analysis of DNA better than algorithmic analysis? Which one is more accurate? What do you think is wrong with algorithmic analysis?
   4. Is the FBI using models to generate DNA matches or it is done in a lab by an examiner?

---------------------------------------------------------------------------------------------------------------------

4. A list of expert sources (at least 3) who can offer subject matter expertise and/or a broader theoretical, practical, formal or ethical perspective on the system you are analyzing. These should be individuals with *specific* interest and background in the system you're looking at, not just general "AI" or "algorithms". List at least 3 questions you would ask of each.

We’ve already reached out to a few expert sources and two of them have accepted our interview request. We’ll be talking to them next week.

1. Brett Williams, CEO of GEDmatch [Interview Confirmed]
   1. How did GEDmatch begin? Where did the idea for this service come from?
   2. How do you handle law enforcement that has been using your platform outside of its intended use?
   3. When sites like 23andMe and AncestryDNA update their data and their results change, does GEDmatch also update accordingly?
   4. Does this have any impact on the matches that people find using your service?
   5. Does your service look at similar genetic markers as ancestry testing services or does it fit better into another type of DNA testing like forensic or medical? How does it differ from other testing services?
2. Brooklyn DNA Testing Lab [Interview Confirmed]
   1. How does testing differ based on what the test is being used for? (ex: ancestry, forensic, medical, paternity, etc..)
   2. Do you think that law enforcement should be using ancestral DNA test results in place of forensic?
   3. Are there any dangers associated with this?
   4. Looking to the future, should people be concerned about how their DNA will be used?
3. Ankur Dalia, Associate Professor of Biology at Indiana University [Awaiting Confirmation]
   1. Could you tell us more about your work in genome engineering and editing?
   2. What are some flaws in consumer DNA testing systems?
   3. What impact can these flaws have?
   4. How far away are we from “designer babies”? What are the ethical implications behind them?
   5. Do you see a “Gattaca” scenario happening where non-designer babies become second-class citizens?