AGENDA

AI/ML

29th November 2023 at 10am EST

1.0	Welcome ● Introductions ● Discussion of group's history ● ·Discussion of subgroup leadership	*Beatrice Kaiser*	*10 min*
2.0	Discussions ● Conversational AI ● Digital attribution ● Definitions/guidance around streamlining data access.	*Everyone*	*40 min*
3.0	Next Steps	*Everyone*	*10 min*

Zoom Details:

Join Zoom Meeting
https://us02web.zoom.us/j/85028744504?pwd=Sm5KYlIYdUdwM3dwajlKNGxNK2JUQT09

Meeting ID: 850 2874 4504
Passcode: 811299

Attendees:

1. Beatrice Kaiser (BK)
2. Gemma Brown (GB)
3. Jessica Seegobin (JS)
4. Ma'n H. Zawati (MZ)
5. Rosalyn Ryan (RR)

Minutes:

BK: Following introductions from attendees, BK introduced the group:

- The history of the discussions around AI/ML was initiated at the last roadmap in 2021/22 and the discussion was around bias in machine learning and algorithms, it was initiated by Tina Hernandez-Boussard, and there were overlap with this group and the Diversity in Datasets group, and because of the overlap, there was a decision to pause. Just because the discussions were synonymous, they didn't want to duplicate work, and now that work has been done, there are a lot more questions about AI, so it's a prime time to reinvigorate the group and have gotten lots of interest.

- In terms of leadership, Yann has spoken to both about being co-leads, and Tina is still going to be involved. BK will connect the three after this meeting.

BK: In terms of moving forward, there isn't a specific focus needed, which is a good starting point for bringing people in.

Spoke about GA4GH subgroup leadership:

- Being a subgroup lead includes people who are engaged, expertise and interested in policy development. Involvement is dependent on the amount of time and work you want to contribute. For e.g., there are leads that write blog posts/briefs and disseminate in the group, and others lead discussion, and the substantive work will be passed onto policy analysts, who write the substantive document and send it to the leads for review. So, there isn't a set of rules, but leads are expected to be involved in setting agendas, leading meetings, etc.

- In terms of co-leads, there will likely be 3 on this project. Generally, at REWS, there is at least one policy analyst that's assigned for the subgroup; for this one, it'll likely be GB.

Suggests starting with a scoping literature review to make sure there's no duplicating of information. Before getting into policy writing is this first scoping literature review; after that, there is the development of a product (blog, policy) and then further down the line, there are steps to approve the policy, but those details won't be addressed in this meeting.

MZ: Agrees it's an exciting, vast topic. Expressed seeing GA4GH as being anticipatory as possible, and not reactive. One of the topics of discussion is conversational AI, the ability to combine natural language processing with machine learning. This can be used to query datasets, help researchers who have received access data to query the data they've received. In the field of data sharing, MZ's involved with groups and acknowledges there's risks and suggests GA4GH should think of these tools and work out the core components.

Expressed that it's been hard to streamline processes for this. One of the issues is that data access is not looked at from an administrative process but only the application is reviewed quickly; doesn't mean that access is given. After that, the whole data access agreement process is heavy in time (average 3-6 months depending on project), and then the transfer of data.

There's lots of questions/queries that need to be dealt with by access offices. So, the researcher can query and get responses in terms of the nature of data but there are limits to that (in terms of identification). The same thing with data that has been collected under a certain agreement, where there should be a system in place that respects the core elements of the access agreement.

GB: Suggests the balance between promoting and supporting researchers accessing and maintaining patients and participants. General trust is slightly wavering.

MZ: Agreed, it's an important point. MZ has been an access officer for the International Cancer Genome Consortium and Human Cell Atlas Program; they're always trying to find ways to streamline and keep that trust. Transparency is key but when thinking about AI, hopefully this is a basis for trust. Part of the way to make conversational AI is to think of frequent questions from users, and respect for what people have consented for.

The other thing is one of access. You want to create a tool that's open-sourced so people can implement this in their own project but if it doesn't have what we hope to be a GA4GH seal, what's needed is to consider parameters for consent and for agreement. Hopefully, something like this can be a building mechanism, esp. if it's implemented at the right times and specific controls. MZ has also spoken to Mark F. who's also interested in being part of the subgroup.

Other things that come to mind is the concept of digital attribution. During COVID, one of the challenges was having access to open datasets quickly. The idea was in this process to think of when someone is querying datasets, this automatically gets embedded in the data to ensure the attribution is always there and it doesn't have to make processes that make it harder for people to have access.

RR: Data access portion is a step down the road. Warned that there's a data access committee that meets, and there's a need to make sure there's no overlap in terms of data access and REWS. Would like to look at not necessarily accessing data from an AI/ML perspective but defining what that is, how it's developed from a use case perspective. From the world of policy makers with genomic information and patients getting tested, they're not at the point of understanding how AI is put together. If testing of datasets is not properly validated, the results of the database would be the same just as in the non-AI world.

Proposed questioning how do we prove AI will not cause harm and there won't be artificial responses because the dataset hasn't been checked? Proposes to go back a step before and ask, "what is AI? How is it formulated? How is machine learning applied and what are the rules/restrictions? What's the government's role?"

Indicates that formatting is also important; how fields line up? In combing through data that's not built properly, the only results that will be obtained are 'very truthful' or 'not truthful'. Suggests getting some definitions on what it's going to look like, what things need to be addressed in terms of policy, bias.

GB: Asks for clarity from RR whether she's suggesting developing some guidance around testing and validation? Suggests guidance on using AI to streamline the data access process.

MZ: Responds that AI is very vast. There are common elements – trust, respect, transparency, and concerns around datasets being harmonized. Asks RR her thoughts about how AI can help curate and harmonize these datasets.

RR: Responds that AI is a tool and way of looking at information in a faster, more efficient way. This is how people on the provider side define it. It is not a finished product. That said, if it's going to provide answers by combing through data to get to those answers - before applying a tool to the data, we're down to the basic questions all researchers have about validity and consistency. Suggests to question how we get these tools to verify that the right thing is being done?

MZ: Part of this discussion needs to include some type of validity. You can't train a tool to answer the wrong questions. The training needs to be quick. In terms of the answer you would get, you need to make sure answers are well-informed. Moreover, there's a need to make sure that the dataset is well collected and harmonized, this is a necessity with or without AI.

RR: Agrees that MZ's approach focuses on access to datasets before we know that AI can do the job in combing through, and needs to make sure that AI is trained on the right tools to identify what is needed in the query.

MZ: The idea behind this is to work with colleagues not only from the ethical and legal world but also AI as part of a working group. There are things that come with processes where we can have useful inputs that can also become a template for the future.

RR: Says that she comes from a technology background. That's where it's a combination of what providers need and the questions technology has to answer.

MZ: Agrees that this can sometimes be subtle.

RR: Agrees it could be about bias and harm

MZ: Believes the first part is to establish what the principles are. Having colleagues that are experts in the technical field is to translate those principles into the language of these tools and put together a policy, checklist or template that lets us have clear definitions.

RR: Believes a checklist is good.

BK: Expresses consensus over addressing the governance of the creation of AI tools.

MZ: Suggests that governance is a big term. What's needed is to take this idea of conversational AI and to work through all of its processes and this has a major IT component. It has policy, compliance, respect for persons elements also. It's useful to think of something that's concrete. For people already working on this,

it'll help them to think about other important elements and be included in the design of the tool, or else this falls into issues of harm.

RR: Agrees with the point. It's critical to identify issues between conversational AI that looks for word patterns and info in a database versus creating a tool that's used to improve healthcare. We must be careful that what we're doing is based on how AI is used in a healthcare environment, compared to a research environment. The value of AI right now is being able to take a task that is causing provider concern and streamline it. The datasets have to be trained to do things that eliminate the burden of combing through vast amounts of data. There's a need to be specific if we're talking about conversational AI, compared to a technical medical process.

RR: Asks for links to be sent that were posted in the chat.

BK: Asks GB to email the link directly. Says will be in touch with Tina.

GB (from chat): https://www.ft.com/content/f8f5184c-0945-408d-8bcf-c101fa801dec

Next Steps:

BK: Start with scoping review. Suggests taking this conversation on our side (GB, BK), and sending an email to recap. Then, early next year, start with subgroup meetings. Suggests using this to develop some starting general questions for the group that can help direct and assess what other people can bring to the plan. GB and BK can come back with a list of what's relevant and use that to direct, and ask MZ and RR to share resources that they're already thinking of.