How can cross-sector collaborations drive responsible use of AI for genomic innovation?

Behind the Genes Transcript

**Francisco: Welcome to Behind the Genes.**

[Music plays]

**Rich:** The key is to deliver what we see at the heart of our mission which is bringing the potential of genomic healthcare to everyone. We can only do that by working in partnership. We bring our expertise and those unique capabilities. It’s about finding it in different ways, in different collaborations, that multiplier effect, and it’s really exciting. And I think the phase we’re in at the moment in terms of the use of AI in genomics is we’re still really early in that learning curve.

[Music plays]

**Francisco: My name is Francisco Azuaje, and I am Director of Bioinformatics at Genomics England. On today’s episode I am joined by Karim Beguir, CEO of InstaDeep, a pioneering AI company, Harry Farmer, Senior Researcher at the Ada Lovelace Institute, and Rich Scott, CEO of Genomics England. Today we will explore how Genomics England is collaborating with InstaDeep to harness the power of AI in genomic research. We will also dive into the critical role of ethical considerations in the development and application of AI technologies for healthcare. If you’ve enjoyed today’s episode, please like, share on wherever you listen to your podcasts.**

[Music plays]

**Let’s meet our guests.**

**Karim:** Hi Francisco, it’s a pleasure to be here. I am the Co-Founder and CEO of InstaDeep and the AI arm of BioNTech Group, and I’m also an AI Researcher.

**Harry:** I’m Harry Farmer, I’m a Senior Researcher at the Ada Lovelace Institute, which is a think-tank that works on the ethical and the societal implications of AI, data and other emerging digital technologies, and it’s a pleasure to be here.

**Rich:** Hi, it’s great to be here with such a great panel. I’m Rich Scott, I’m the CEO of Genomics England.

**Francisco: Thank you all for joining us. I am excited to explore this intersection of AI and genomics with all of you. To our listeners, if you wish to hear more about AI in genomics, listen to our previous podcast episode, ‘Can Artificial Intelligence Accelerate the Impact of Genomics’, which is linked in this podcast description. Let’s set the stage with what is happening right now, Rich, there have been lots of exciting advances in AI and biomedical research but in genomics it’s far more than just hype, can you walk us through some examples of how AI is actually impacting genomic healthcare research?**

**Rich:** Yeah, so, as you say, Francisco, it is a lot more than hype and it’s really exciting. I’d also say that we’re just at the beginning of a real wave of change that’s coming. So while AI is already happening today and driving our thinking, really we’re at the beginning of a process. So when you think about how genomics could impact healthcare and people’s health in general, what we’re thinking about is genomics potentially playing a routine part in up to half of all healthcare encounters, we think, based on the sorts of differences it could make in different parts of our lives and our health journey. There are so many different areas where AI, we expect, will help us on that journey. So thinking about, for example, how we speed up the interpretation of genetic information through to its use and the simple presentation of how to use that in life, in routine healthcare, through to discovery of new biomarkers or classification that might help us identify the best treatment for people.

Where it’s making a difference already today is actually all of those different points. So, for example, there’s some really exciting work we’re doing jointly with Karim and team looking at how we might use classification of the DNA sequence of tumours to help identify what type of tumour - a tumour that we don’t know where it’s come from, so what we call a ‘cancer of unknown primary’ - to help in that classification process. We’re also working with various different people who are interested in classification for treatment and trials, but there’s also lots in between recognising patterns of genomic data together with other complex data. So we’ve been doing a lot of work bringing image data together with genomic data and other health data so that you can begin to recognise patterns that we couldn’t even dream of. Doing that hand in hand with thinking about what patients and participants want and expect, how their data is used and how their information is held, bringing it all together and understanding how this works, the evidence that we need before we can decide that a particular approach is one that policymakers, people in healthcare want to use, is all part of the conversation.

**Francisco: Thank you, Rich, for speaking of cutting-edge AI applications and InstaDeep. Karim, could you give us a glimpse into your work and particularly how your technologies are tackling some of the biggest challenges in genomic research?**

**Karim:** Absolutely, and I think what’s exciting is we’ve heard from Rich and, you know, this is like the genomics expertise angle of things and I come from the AI world and so do most of the InstaDeep team. And really what’s fascinating is this intersection that is being extremely productive at the moment where technologies that have been developed for like multiple AI applications turn out to be extremely useful in understanding genomic sequences.

This is a little bit, our journey, Francisco. Back in 2021/2022 we started working on the very intriguing question at the time of could we actually understand better genomic sequences with the emerging technologies of NLP, natural language processing. And you have to put this in context, this was before even the word ‘generative AI’ was coined, this was before ChatGPT, but we had sort of like an intuition that there was a lot of value in deploying this technology. And so my team, sort of like a team of passionate experts in research and engineering of AI, we tackled this problem and started working on it and the result of this work was our nucleotide transformer model which we have open sourced today; it’s one of the most downloaded, most popular models in genomics. And what’s interesting is we observed that simply using the technologies of what we call ‘self-supervised learning’ or ‘unsupervised learning’ could actually help us unlock a lot of patterns.

As we know, most of genomics information is poorly understood and this is a way actually, with using the AI tool, to get some sense of the structure that’s there.

So how do we do this? We basically mask a few aspects of the sequence and we ask the system to figure them out. And so this is exactly how you teach a system to learn English, you know, you are teaching it to understand the language of genomics, and, incredibly, this approach when done at scale - and we train a lot on the NVIDIA Cambridge-1 supercomputer – allows you to have results and performances that are matching multiple specialised models. So until then genomics and use of machine learning for genomics was for a particular task, I would have developed a specific model using mostly supervised learning, which is, I am showing you a few examples, and then channelled these examples and tried to match that, and so essentially you had one model per task. What’s really revolutionary in this new paradigm of AI is that you have a single model trained at very largescale, the AI starts to understand the patterns, and this means that very concretely we can work with our partners to uncover fascinating relationships that were previously poorly understood. And so there is a wealth of potential that we are exploring together and it’s a very exciting time.

**Francisco: What you’re describing really highlights both the potential and the opportunities but also the responsibility we have with these powerful tools, its power, and this brings up some important ethical considerations. And we have Harry… Harry, we have talked about ethics frameworks in research for decades but AI seems to be rewriting the rulebook. For your work at the Ada Lovelace Institute what makes AI fundamentally different from previous technologies when it comes to ethical considerations and how does this reshape our approach to ensuring these powerful tools benefit society as a whole?**

**Harry:** So I think when you are considering these sorts of ethical questions and these sorts of ethical challenges posed by AI and genomics it really depends on the sort of deployment that you’re looking at. From the conversation we’ve had so far, I think what’s been hinted at is some of the diversity of applications that you might be using AI for within the context of genomics and healthcare. So I think there’s obviously big advances that have been alluded to in things like drug discovery, in things like cancer and cancer diagnosis, also these advances around gene editing, all of which have been on steroids, by artificial intelligence and particularly machine learning and deep learning.

The area that we have been looking at at the Ada Lovelace Institute, and this was a project that we were doing in collaboration with the NCOB, the Nuffield Council on Bioethics, was looking at what we were calling ‘AI-powered genomic health prediction’, which is very related to a technique called ‘polygenic scoring’, for those who might be interested. And that’s looking at the emerging ability to make predictions about people’s future health on the basis of their DNA, and it was thinking about what that ability might mean for UK society and also for how we are thinking about and delivering healthcare in the UK.

Now, thinking about what the ethical challenges might be for that, I think you need to think about what specifically AI is bringing to that technique, so what it’s bringing to genomic health prediction. I think with some of the other deployments, the list of things that AI is bringing is quite similar, so it’s helping with data collection and processing, so speeding up and automating data collection and preparation processes that otherwise are quite slow and very labour-intensive. AI’s also helping with the analysis of genomic and phenotype data, so helping us to understand the associations between different genomic variations and between observable traits, and this is something which without AI can often be prohibitively complex to do, and it’s also sometimes suggested that on the deployment end AI can be a tool that can help us use genomic insight in healthcare more widely. So one example of this might be using an AI chat bot to explain to a patient the results of a genomic test. That’s something that’s only been mooted and I don’t think there are current examples of that at the moment but that’s one of the downstream applications of AI in the context of genomics.

So in terms of what AI’s actually doing and what it’s bringing, it’s really just making possible things that we’ve been trying to do in genomics for some time, making these things easier and cheaper and in some cases viable. So really it’s best to see it as an accelerant for genomic science; it doesn’t present any brand-new ethical problems, instead what it’s doing is taking some fairly old ethical challenges and making these things far more urgent. So in terms of what those problems actually are, some of the big ones will be around privacy and surveillance, genomic health predictions produce a lot of intimate sensitive data about people and generating those insights requires the collection and the storage and the processing of a lot of very sensitive data as well. We also have issues related to privacy around genomic discrimination, so this is the worry that people will be treated differently and in some cases unfairly on the basis of health predictions made about them. And one of the really typical examples here is the worry that people might face higher insurance costs if they’re found through genomic testing to be more likely to develop particular diseases over their life course.

And then you also have a bunch of issues and questions which are more structural, so these are questions about how the availability of this kind of insight into people’s future health might change or put pressure on existing ways of thinking about health and thinking about healthcare and some extreme cases thinking about the social contract. So these are questions like does the viability of genomic health prediction lead to a radically more preventative approach to healthcare and what might this mean for what the state demands of you as a user of healthcare and as a recipient of that. And there are also some important questions about the practicalities of delivering genomic medicine in the NHS, so questions like how does the NHS retain control and sovereignty over genomic analysis and data capacities, how do we test their efficacy at a public health level, and also – and this is something that we might talk about a bit later – what’s the best deployment model for these capacities. So that’s some of the ethical and I think policy challenges that we need to be dealing with in this space.

**Francisco: Thank you, Harry. And those principles you have outlined provide a solid foundation for discussing different types of applications.**

[Music plays]

**Let’s talk about the InstaDeep and Genomics England partnership that is investigating the application of InstaDeep’s powerful foundation model, the nucleotide transformer, and other cutting edge techniques to address several challenges in cancer research. I have the privilege of working closely with this partnership and the potential here is immense. Karim, could you break down for our listeners what you are working on together and what innovations you are aiming for?**

**Karim:** Absolutely, Francisco. Actually, we are very excited by the collaboration with Genomics England. Genomics England not only has one of the best data assets in the world when it comes to genomics, like a very well curated dataset but also a wealth of expertise on these topics, and on my side the InstaDeep team brings fundamental knowhow of machine learning models but also, as you mentioned, like powerful developed models already, such as our nucleotide transformer and others.

The culture of InstaDeep has always been to build AI that benefits everyone – this is literally in our mission – and so in particular, specifically on like current topics, really like the goal is to try to identify partners between genomic sequences of patients and the particular phenotypes or approaches. And one of the key projects, which I mentioned that, is the one of cancer of unknown primary origin. So when you have situations where you are not sure where a particular cancer emerged from it is critical to be able to extract this information to have the best potential care, and this is actually something where understanding of genomic sequences can bring this capability. And so we’ve been getting some successful results in the collaboration but in many ways this is just the beginning. What we are seeing is a great wealth of possibilities linking genotypes, so the information which is on the sequences themselves, the genomic sequences, and phenotypes, like the particular state of the patient, and the fact that the Genomics England team has those joint datasets creates incredible opportunities. So we are looking at this really like identifying together what are the most useful ‘low-hanging fruits’, if you want, in terms of like potentially improving a patient’s care and moving forward from that.

**Francisco: And this collaborative approach you are describing raises questions about accelerating innovation in general. When two organisations like Genomics England and InstaDeep come together it’s like a multiplier effect in terms of expertise, data, and other resources. Could you both share how this partnership is accelerating discoveries that might have taken years?**

**Rich:** Yeah, I mean, I think this… Francisco, you frame it really nicely because this is what makes it so exciting to be in our position at Genomics England because what we do is we bring the particular understanding and expertise, digital infrastructure and custodianship of the National Genomic Research Library together, but actually the key is bringing the potential of genomic healthcare to everyone. We can only do that by working in partnership, we bring our expertise and those capabilities. And, as you say, it’s about finding it in different ways, in different collaborations, that multiplier effect, and it’s really exciting. And I think the phase we’re in at the moment in terms of the use of AI in genomics is we’re still really early in that learning curve.

And so, as you’ve heard already through what Karim and I have said and also what Harry has said, there are multiple different aspects that we need to look at together, bringing different angles and understandings, and we see ourselves… We often describe ourselves as a ‘data and evidence engine’, that final word ‘evidence’ is really important and it comes in the round. So Harry really eloquently talked about a number of different considerations from an ethical perspective that need to be there. What we need if we’re going to move genomics forwards in terms of its potential to make a difference for people’s lives, we need evidence around clinical efficacy of different approaches, that’s absolutely a given and everyone always jumps at… so it’s almost first in line. We need understanding about the health economics, you know, how much difference does it make for a particular investment, is it worth that investment. Critically, it also is founded on, you know, how you might use this technology in different ways, how you use it in clinical pathways, you know, is it something that actually is addressing the particular questions which really hold back the delivery of better care. Also in that evidence piece is an understanding of patients’ and participants’ expectations on how their data might be used, their expectations on privacy, the expectations that we have on understanding how equitable the use of a particular approach might be, or at least our understanding of how confident we are about the equity of the impact, and it’s bringing together those different perspectives. And that’s one of the things that helps us construct the team at Genomics England so we have the expertise to help others access the data in the National Genomic Research Library for purposes our participants support but also help generate that sort of rounded package of evidence that will end up moving the dial. So that it’s not just about proving a cool widget, because that’s great on its own, what drives Karim and the team is to make a difference in terms of outcomes, and that’s exactly what drives us and our participants too.

**Francisco: And this and other partnership approaches brings up important questions about responsible innovation, and this naturally leads us to the next question for Harry, how do we harness these powerful tools when protecting our communities?**

**Harry:** Yeah, so if we are thinking about over-surveillance and the ways that vulnerable groups might be affected by the use of genomics and healthcare, I think we’re talking about at least two different things here. So one problems around the representativeness of data is it does lead to issues which you could classify as issues of differential accuracy. So in the context of genomic prediction what you have is genomic predictive tools being more accurate for white Europeans and those with white European ancestry compared to other population groups. And this is a product of the fact that genomic datasets and genomic predictions, the terminologies don’t port well between different populations, which means if you train a genomic predictive tool on a bunch of people with white European ancestry the predictions you might make using that tool for other groups won’t be as accurate as for the white Europeans. And this can be actively harmful and dangerous for those in underrepresented groups because you are making predictions about people which just won’t have the accuracy that you would expect in the context that you were deploying it.

And I already mentioned this a bit in my previous answer, you have worries about discrimination, and there are a few different things here. So with some historically marginalised groups and marginalised groups now there are longstanding historical sensitivities about being experimented on, about particular fears about eugenics and about being categorised in particular ways. And it’s worth saying here that there is obviously a racial dimension to this worry but I think there’s also a class dimension, by which I mean you’re far more vulnerable to being categorised unfavourably if you’re poor or if you don’t have a particular kind of status within society. There is also within discrimination the idea that genomics might be used to explain away differences between different groups which in fact have a political or an economic basis. So one example of this was during the COVID-19 pandemic, there were attempts by some commentators to explain away the fact that non-white communities had worse rates of mortality from COVID to try and attribute a genetic or a genomic basis to those differences rather than looking at some of the socioeconomic factors behind that. So those are some worries as well.

Now, when it comes to protecting particular groups I think there are a few things that can be done fairly straightforwardly. So, one is work to improve the diversity and the representativeness of datasets. Obviously, that’s easier said than done, though it’s a very clear thing that we can aspire towards and there is good work, I’m aware, that is going on in this space, some of which is being spearheaded by Genomics England, amongst other groups. Another is just being very careful about how the results of population level genomic studies are communicated to avoid giving that impression of explaining away differences between different groups simply as things determined by genomics about which we can do nothing rather than things which have historical or socioeconomic bases. But I also think the broader lesson is that some of these harms and these forms of discrimination are things that could theoretically affect anyone; they’re not just limited to affecting marginalised groups.

Genomic health predicting can produce bases for all of us to be discriminated against, things that have nothing to do with our race, our class, our sex or any other protected characteristic. So I think there has to be thinking about how we establish or sure up more universal protections against genomic discrimination. One thing that we can do here is simply stronger data protection law, and one of the things that we talk about in some of our reports is that how data protection law as it stands could do with being less ambiguous when it comes to how it treats genomic data and phenotype data produced as a result of genomic analysis.

[Music plays]

**Francisco: Harry, you are in a unique position at the Ada Lovelace Institute where you bridge this gap between AI developers, researchers, policymakers and the public. Your recent report on AI in genomics with the Nuffield Council on Bioethics offers an important blueprint for responsible AI innovation in general, so based on this cross-sector perspective, what guiding principles do we need to embrace as we navigate this intersection of AI and genomics?**

**Harry:** So I think in addition to the specific recommendations we set out in the final report of that work - which is called ‘Predicting the Future of Health’ and which you can find on our website and also on the NCOB website – I think one of the biggest messages was the importance of finding a deployment model for genomic health prediction that respects that technology’s strengths, what it can actually do, because there are limitations to this technology, and also which avoids circumstances in which the associated risks are difficult to deal with. So another way of putting this is that we need a deployment model that, as well as making sure that we’re ready to cope with the risks of genomic health prediction, the things like law, regulation and governance also proactively tries to design out some of those risks and finds ways of deploying this technology such that those risks don’t present themselves in either as extreme a manner or don’t present themselves in ways which makes them difficult to deal with.

So one question that we posed in our research was whether some ways of integrating genomic health prediction may present more challenges regarding privacy, discrimination and then these other challenges that we’d identified around dependency and fragility and others. And having looked at some of the different broad approaches to using genomic health prediction within the NHS and within the UK’s health system, we found that one presented by far fewest of the risks identified above, while still presenting some of the most certain benefits of genomic health prediction. And this was using it really primarily as a targeted diagnostic tool - and this is a vision in which the NHS uses genomic health prediction quite sparingly in the first instance - and in situations to improve treatment and outcomes for those who are seriously ill or who have been identified as needing to take particular precautions regarding their health. We think the more situational vision has a few advantages. So one, is it allows patient and people using the health service to retain greater control over data. We think that can also have a positive knock-on effect for worries about discrimination. And here what you have is the absence of those pressures to share your data. It means that it’s easier for you as the user of the healthcare system to resist genomic discrimination simply by keeping your data private. And there are some cases where that option… it shouldn’t be the only option but where that option is really important.

And then also one of the features of this vision is that the smaller scale of the use of genomic health prediction, presumed, can make outsourcing to third parties, which the NHS is probably likely to need to do in some cases. It’s also a vision, I think, that overall allows you to capture some of the more certain benefits to genomic health prediction which are about improvements to accuracy in predictions about people’s future health at the margin, and therefore this is a deployment of this technology which is deploying it principally to people who will benefit and we know will benefit from marginal improvements in accuracy to predictions made about their future health rather than wanting to deploy those marginal improvements to the vast majority of the population where the benefit is less certain. So this is a vision we hope sets out a way of getting some of the more certain benefits of this technology while minimising some of those broader more systemic risks.

**Francisco: Thank you, Harry. Karim?**

**Karim:** Totally agree with Harry about the need for smart regulation in the field so that we make sure we have good uses of the technology but avoid the potential pitfalls. I wanted to emphasise two points which I believe are important. First, we are really in a fast-moving situation when we look at like AI progress. We have seen incredible improvements over the last ten years and in particular what we call ‘artificial general intelligence’, which is essentially systems that are matching human cognitive abilities, are now around the corner. This might sound surprising but literally the last obstacles to reach AGI are being solved right now, and this means that in the next 12-24 months you will have systems that are incredibly capable. So this emphasises the need for the type of measures and type of smart approach that Harry has described. And I would say when you look at the intersection of AI and genomics this is a particularly important one and why it’s the case, because so far in genomics our obstacle has not been data, it has been interpretation of a flood of data. The progress that AI is making, like I just described now, means that very soon extraordinary capabilities will be available to improve patients’ outcomes. I want to inject a sense of how important is our conversation today, given what is happening, an exponential progress in AI, exponentially growing data in genomics and relatively exponential potential to build the technology for good. But, like in other fields, we see that AI is an extremely powerful technology and we need to make sure it is used for good in fact and this is why the conversation that we have today is so important.

**Harry:** Obviously I agree with the conclusion to all of this, is that we need to think very hard about the way that artificial intelligence and its deployment in healthcare and also just in many different walks of life is going to be affecting the way we think about public service delivery, affecting the way that we think about scientific development. It’s worth noting, though, that I think one of the biggest challenges from a policy perspective on artificial intelligence is being able to distinguish the wheat from the chaff. There are obviously areas where AI has made huge and incredibly impressive progress over the past few years and where we reasonably expect that to continue over the next few years, but there are also areas where some of the stories being told about the capabilities of future systems probably won’t be matched by the reality, but there is I think a really big and very live debate about exactly what we can reasonably expect from these technologies and therefore what the deployments of them are.

**Francisco: Thank you. We are approaching the end of the episode and I’d like to conclude with a couple of questions. Genomics England has built quite an ecosystem of industry partnerships, how do collaborations like the one with InstaDeep fit into your broader mission for the company?**

**Rich:** So linking this to the conversation that we’ve just been having, which is AI is making a real difference in terms of technologies that we can test, we can develop evidence on, and that is rightly creating excitement, I think our approach… The expectation of our participants is that our role is to sit there and help people develop evidence and you can make judgments on policy based on those and that is what will drive adoption. I think the thing that really excites me for the UK, most particularly in genomics, is our ability to be the place in the world where you can come with a new technology, whether it’s genomic sequencing technology, whether it’s a genomic AI approach to train that to develop evidence on its efficacy, and, if it’s proven to be effective to be worth the bang for the buck to perform to the expectations that patients, the public, would have of it in terms of equity and so forth also to deploy it. I think there is a real reason for excitement around that and it’s a real opportunity that the government has highlighted and that we absolutely buy into that the UK can be the best place to do that for academics and for industry. And our participants see real opportunity and are eager for that work to be done so that we have the evidence on which to decide what should be deployed and where. We see opportunities in all sorts of different areas, so certainly in terms of drug discovery and all the way through to simplifying tasks which at the moment just limit the rate at which the existing uses of genomics in healthcare can happen.

So I think there’s opportunities across the whole length, if you like, the sort of end to end, and the breadth of opportunity, and industry, companies like InstaDeep and others that we work with, are really crucial to that. And what we do is think about the digital infrastructure we need to, you know, have those teams able to interact with within the National Genomic Research Library carrying out their approved research projects. Also what support they need, and that comes in different shapes and sizes, depending on the ask and also the company. So sometimes sort of leaning in more, particularly at the start of programmes, to help people shape the question, working with our participants, thinking about the wider evidence that you might need, for example, those sort of things that Harry’s touched on, but also thinking about what hands-on support companies need, because not every company is anywhere close to Karim and InstaDeep’s expertise. Sometimes this is also about supporting people to have some of those tools that they don’t have or some of the knowhow that’s very specific to areas of genomics, so it’s absolutely crucial to it. And I think that point of the UK being the place to come and develop that evidence in its full breadth so that policy decisions can be made not based on hype but on evidence in the round, on what will make a difference.

**Francisco: And, Karim, looking ahead, also in retrospect, what have been your key learnings about making this cross-sector partnership work?**

**Karim:** We live in an extraordinary time and I want to emphasise the potential of scientific discovery in the next two or three years. AI is going to move from, let’s say, digital style, you know, technologies like coding and maths towards more like science and biology. In particular, genomics is going to be a fascinating area in terms of potential, and I agree with Rich and Harry, it’s all in the end about proving on the ground the potential of those capabilities. And at InstaDeep we are passionate about the tech – I think you might have felt that – but we’re also passionate about the applications. The best results come when you bring expertise from multiple domains; machine learning and AI experts will require the expertise of genomic experts, biologists, healthcare practitioners, to be able to translate the potential of those technologies in concrete outcomes. And we’ve seen this on multiple successful projects we’ve done with Genomics England but really this suggests that we are going to have in the next 3-5 years way more progress than we had in the last five and really my wish is that collectively we seize this opportunity and we do it in a responsible and thoughtful manner.

[Music plays]

**Francisco: We’ll wrap up there. Thank you to our guests, Karim Beguir, Harry Farmer and Rich Scott, for joining me today as we discuss the role of AI in genomics research. If you wish to hear more like this, please subscribe to Behind the Genes on your favourite podcast app. Thank you for listening. I have been your host, Francisco Azuaje. This podcast was edited by Bill Griffin at Ventoux Digital and produced by Naimah Callachand.**

[Music plays]