# Building Trustworthy AI: The Role of Patient and Public Involvement in Healthcare AI Development

**Soumya Banerjee**
University of Cambridge
Cambridge, United Kingdom
sb2333@cam.ac.uk

## Abstract

AI is helping researchers make great strides in healthcare. However, there is a trust deficit in AI when applied to critical areas like healthcare. Hence communicating the beneficial medical applications of AI and engaging the public with healthcare AI research is critical. One way to achieve this is by getting the community involved in co-designing better AI systems in healthcare projects. We argue that AI algorithms for healthcare should be co-designed with patients and healthcare workers, so that they are useful and trustworthy.

We suggest a roadmap for this collaborative approach in AI model building. This will involve actively including patients with lived experience of a disease, as well as creating a research advisory group to walk patients through the process of AI model building. We suggest formulating and scoping a problem, and then generating a hypothesis that patients and scientists agree on.

The road to building trustworthy AI systems may become easier if all stakeholders are involved in co-creating AI models.

## Introduction

Artificial intelligence (AI) is helping researchers and clinicians make great strides in healthcare. Some people, however, believe that AI presents more drawbacks than benefits. Hence communicating its beneficial medical applications and engaging the public with healthcare research is critical.

One way to achieve this is by getting the community involved in co-designing better AI systems in healthcare projects (Banerjee et al. 2022). We believe that integrating patient and public involvement (PPI) in AI projects may help in adoption and acceptance of these technologies.

## Building Better AI Systems

Explainable AI is a set of methods that allow users to comprehend and trust the results and outputs created by AI algorithms. What explanations are useful will be different for computer scientists, clinicians and patients alike. One way to identify which explanation is most useful is to let people co-develop models, play with these models and generate their own explanations.



We argue that AI algorithms for healthcare should be co-designed with patients and healthcare workers, so that they are useful and trustworthy.

In our work, a patient with a severe mental illness approached us. The patient had bipolar disorder and was taking a medication called lithium: a very effective medication, but which can damage the kidney if taken for many decades. The patient asked us whether we could use AI to understand whether stopping lithium may reverse kidney damage. We analysed hospital data to find patients who had stopped lithium (after taking it for many years), and if this had reversed their kidney damage.

Having a patient approach us with a clear hypothesis to set the research agenda was profound. This patient-led approach prevents study participants from becoming mere datapoints and potentially becoming marginalised in the research process.

We built simple models to get the process started. Initially, we built simple linear regression models and explained them to the patient. Subsequently, we added layers of complexity to these models and in monthly meetings, took steps to explain and validate them with everyone involved.

We suggest a roadmap for this collaborative approach in AI model building. This will involve actively including patients with lived experience of a disease, as well as creating a research advisory group (RAG) to walk patients through the process of AI model building.

We suggest formulating and scoping a problem, and then generating a hypothesis that patients and scientists agree on. Data scientists should then start by building simple models and explain them to patients and clinicians. In our work we built simple linear regression models which can be explained very easily. This will lay the groundwork for progressively building more complex models and explaining each step of the model building process to patients.

We specifically suggest: 1) including patients with lived experience of the disease and carers, 2) creating a Research Advisory Group (RAG) and using these group meetings to involve patients and carers in all stages of the scientific process (starting from hypothesis generation). We also recommend explaining the process of AI model building, starting with simple (e.g. linear) models. We suggest using freely available AI models that run in the browser (such as the Teachable Machine from Google) to explain the basics of

AI to patients. These meetings should be repeated to elicit feedback from the stakeholders, explain model predictions and get guidance on model modifications.

## The Benefits of Building Together

Perhaps the most important benefit of this methodology is that co-developing models transparently, and marrying the technical with the human, may help people build trust in AI systems. It enables patients to not only help shape the models, but also develop a relationship of trust with the team of researchers behind them, because they listen to people and try to understand their needs and allay their concerns.

Opening AI models to investigation not only enables rich interaction, but pinpoints errors and areas of improvement too. It can also highlight unconscious biases.

Data scientists, clinicians and patients worked closely with each other to develop a data science solution that is of use in a completely different domain (serious mental illness). This imbibes principles of affective human-centered computing. It underscores how data scientists can build AI tools that can be used for good in domains different from computer science (Regli 2017).

Our work is similar to user-centric participatory design (He et al. 2019) and affective computing (Calvo et al. 2014). We used patient public involvement in a user-centric participatory design loop. Our study was patient led and initiated, which has many advantages (Ball et al. 2019).

We explained how AI will be used on clinical data and how the expected outcomes might benefit patients. In turn, we learned from patients and carers about important features of the data, and about the concerns that must be addressed to implement AI models in practice - including the potential for inadvertent discrimination by AI (Health 2019).

## Caveats and Limitations

Unfortunately, such a participatory approach does not happen often because it can be difficult. It is much easier to get access to data and write papers than overcome logistical challenges like recruiting patients and finding time for everyone to meet in a Research Advisory Group. There are also few incentives for researchers to work in this way as we typically gain tenure for publishing impactful papers rapidly, rather than experimenting with new engagement techniques.

We believe scientists need more training in engaging effectively with stakeholders to facilitate successful collaborations and co-design in AI. We would love to see the incentive structure modified so we could start to see ripple effects that would benefit patients and researchers alike.

We would like to see university curricula that enable data scientists to better engage with different stakeholders and end-users of AI. At the very least this curriculum should give data scientists a moral compass, like doctors. It should imbue in them a sense of moral responsibility for the AI systems that they create and a need to involve all stakeholders and end-users (like patients). If scientists were better trained to engage with the public and end-users, the process of building trustworthy AI systems may become easier.

We are also getting to an age where data scientists have disproportionate power, so anything to guide them would be very helpful. This may take the form of teaching how to collaborate with end-users (like patients) or courses in ethics.

## A Collaborative Future

Our arguments also apply to domains other than healthcare. AI is increasingly being integrated into public decision making. For example, authorities in Netherlands deployed a benefits fraud detection AI system and falsely accused people who are ethnic minorities or have low income. People who are affected by decisions made by AI systems have a right to be involved in understanding these systems. If possible, they should also have a say in how these systems are designed. We need to involve citizens in understanding how these AI systems work and give them a voice.

A collaborative process ideally necessitates the creation of a research advisory group to involve all stakeholders in the process of building AI models. We believe this is how all AI research in healthcare and potentially other critical domains should be conducted: where everyone, including end users (patients) and domain experts (like clinicians and data scientists) can work together effectively with a shared goal, and collectively benefit from the results.

An understanding of what AI can and cannot do will build trust and allay fears. A realistic appraisal of risks and benefits may help in adoption and democratise access to AI for healthcare, which is in everyone's interests.

We have difficulty trusting what we do not understand. We may easily trust what we create (and hence understand). The road to building trustworthy AI systems may become easier if all stakeholders are involved in co-creating AI models.

The word AI has become associated with dreadful things like machines taking over the world. One way to change that narrative is to give people the power to see what they can do with AI and let them generate their own understanding of these systems.

Involving people in all aspects of AI development and deployment can potentially help *humanize* AI. If AI is to have a bright future in society, we need to involve people in understanding and building AI.

## References


Ball, S.; Harshfield, A.; Carpenter, A.; Bertscher, A.; and Marjanovic, S. 2019. Patient and public involvement in research: Enabling meaningful contributions.

Banerjee, S.; Alsop, P.; Jones, L.; and Cardinal, R. N. 2022. Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies. *Patterns*, 3: 100506.

Calvo, R. A.; D'Mello, S.; Gratch, J.; and Kappas, A. 2014. *The Oxford Handbook of Affective Computing*. Oxford University Press.

He, X.; Zhang, R.; Rizvi, R.; Vasilakes, J.; Yang, X.; Guo, Y.; He, Z.; Prosperi, M.; Huo, J.; Alpert, J.; and Bian, J. 2019. ALOHA: developing an interactive graph-based visualization for dietary supplement knowledge graph through

user-centered design. *BMC Medical Informatics and Decision Making*, 19: 150.

Health, T. L. D. 2019. There is no such thing as race in health-care algorithms. *The Lancet Digital Health*, 1: e375.

Regli, W. 2017. Wanted: Toolsmiths. *Communications of the ACM*, 60: 26–28.