BenevolentAI

AI Governance & Risk Management Model

Who we are

BenevolentAI Group is an organisation that creates and applies artificial intelligence (AI) technologies to transform the way medicines are discovered and developed.

We seek to improve patients' lives by applying technology designed to generate better data decision-making and in doing so lower drug discovery and development costs, decrease failure rates and increase the speed at which medicines are generated. We combine advanced artificial intelligence (AI) and machine learning (ML) with cutting-edge science to decipher complex disease biology and discover optimum therapeutic interventions.

AI for Science and Drug Discovery

In a field like drug discovery where every falsified hypothesis has a financial and human cost, there is a strong motivation to pick the right scientific battles. The most challenging part of the scientific process is, arguably, coming up with a novel yet robust hypothesis in the first place. It takes a combination of collective experience and expertise, individual ingenuity, and perseverance just to be in the position to pose the right question. This point is well-appreciated in the field of mathematics where the name attached to a conjecture — think Fermat, Poincaré, and, no doubt eventually, Riemann – often outlives the solvers and their proof.

In the past decade, this task of scientific hypothesis generation has been transformed by the huge increase in the volume and types of scientific data available. With the revolution of 'omics data to nation-scale biobanks and millions of open-access scientific publications, we are privileged to be in the midst of a parallel revolution in AI that is providing an ever-increasing array of algorithms and computational tools to uncover, at unprecedented speed, the scientific insights hidden in all that information.

At BenevolentAI, we strive to develop and combine advanced ML algorithms with cutting-edge science to decipher complex disease biology and discover optimum therapeutic interventions. The use of AI allows us to overcome the familiar human limitations of expertise silos, cross-domain ignorance, and subject biases that threaten to cancel out the benefits of Big Data. We use AI to profile human diseases – not individuals – to enable significantly faster, preclinical drug development with a higher probability of efficacy and progressability.

We adopt the following four principles of AI development at BenevolentAI:

- Data Privacy and Security,
- Safety,
- Reliability, and
- Transparency.

Data Privacy and Security

At BenevolentAI, we maintain a Privacy by Design (PbD) policy, which is a framework based on proactively embedding privacy and security requirements into the design, development, deployment, and operation of our AI applications, IT systems, networked infrastructures, cloud platforms, business practices, and operations. This ensures that privacy and security become our default mode of operation across the business on deployment.

Security by Design is an approach we deploy that integrates risk-appropriate security controls into all aspects of the design, development, implementation, operation, and decommissioning of applications.

Safety

We observe the following two safety principles.

First, at every stage of the drug discovery and development pipeline, the output of any AI model is **not treated any differently to the output coming solely from traditional, human-driven, methods**. In particular, the safety protocols and obligations, from preclinical models to clinical trials, developed and honed by the pharmaceutical industry and international regulatory agencies over decades are applied and met in full by AI-driven outputs.

Second, we ensure that at every stage **all AI recommendations are assessed by human scientists**. Where the output of one ML model feeds into another ML model further down an analysis pipeline, the intermediate results are assessed for accuracy and quality, meaning that real-world outputs are reviewed by human scientists before being acted upon. For example, we have implemented a human-in-the-loop process where an assessment of the data used by our AI systems to learn about disease biology is often first processed and compiled from the original data sources using ML algorithms before being used in our project pipelines.

Reliability

A Reliable AI/ML model is simply one that performs as intended within user-specified limits. With drug discovery scientists as our primary users, the performance of our AI systems are measured by their ability to accurately and usefully model the biology of human diseases. We focus on the three sequential stages shared by all ML pipelines: Data, Models, and Validation.

Reliable Data

For the purposes of learning to model a system as complex as human biology from data, it is often a case of “more is more”. We make a great effort to assemble the largest data sets across as many relevant data types as possible, at multiple levels of abstraction. This includes structured biomedical databases, unstructured data from scientific pieces of literature, electronic health records, and genomics data sets, gathered across public sources, through our own laboratories, and in partnership with industry and academic collaborators. Given that one significant limiting factor in modern ML models' performance is the size of data sets, we seek to integrate data from multiple contexts and, in some cases, even consider expanding data sets at the expense of introducing noise (e.g. using distant supervision).

Nevertheless, this push for size is counterbalanced by the need to address the systemic biases and noise in biomedical data sets – such as topic bias in the scientific literature, batch effects, and spurious correlations from confounders in molecular data. To address this we employ a variety of quality assurance methods ranging from manual filtering and review to empirical benchmarking to ML-driven data pre-processing. We have dedicated resourcing focused on monitoring and improving data quality throughout our platform, an internal data quality reporting system, and a framework defining clear ownership of scientific and technical data quality assurance as it pertains to specific data resources

Our AI development processes have to consider sources of biases such as the over-representation of certain patient groups, the lack of data about others, and challenges of properly assessing the merits of strong correlations. Patients represented in biomedical data sets tend to be predominantly male and of European descent. This lack of diversity means the medical products that would come from data analysis may not consider all groups in society

We can help remedy this problem in different ways. One viewpoint is “if you can’t measure it, you can’t manage it”. Thus, the first step is to assess the quality of the data with regards to diversity - we have done this across many of the data sets that we use in our products and have shared our insights with the scientific community through blog posts and presentations. This now allows us to actively seek out data sets that fall into those gaps that we discovered through this analysis. However, we recognise that while more diverse data will certainly improve the current status quo, we need to also understand what biases can be alleviated at the point of model development and benchmark definition all the way to the diversity of our teams.

There are two further considerations specific to the data we handle at BenevolentAI. First, it is that scientific knowledge is highly dynamic. We constantly improve, update and deprecate data sets on a regular schedule, while ensuring that the information that feeds into our AI systems is time-stamped and placed in the context of scientific development. Second, we recognise that biological knowledge is highly contextual where tissue, cell-type, and disease specificities are paramount.

Reliable Models

Reliability in ML modelling can be encapsulated in the broad objective of minimising the three *gaps* – the generalisation, domain and reproducibility gaps. The risk here is the burden of ineffective models or, worse, misleadingly effective ones. The tools to reduce and monitor the generalisation gap in supervised learning are universal, such as regularisation, separate training and test data sets and data augmentation strategies. The domain gap, on the other hand, is highly specific to the application under consideration. At BenevolentAI, we wrestle with the domain shifts of different cell-types and tissues, animal models vs. humans and cell-lines vs. patient samples, ensuring that models trained in one domain can be successfully transferred to another. The issue of reproducibility is one of statistical rigour, and it includes the need for the robustness of model predictions to variations in initial conditions, to variable selection with transparent statistical guarantees (e.g. false discovery rates).

Reliable Validation

Because perfectly reliable data and models do not exist, we need reliable validation processes to understand their advantages and shortcomings, both within and outside their domains of applicability. The top requirement here is the alignment of evaluation and usage, where models and the data they consume are evaluated on measures that most closely match their actual use cases. For example, if the objective is to produce a shortlist of candidates for a drug screening assay, then the top-1 prediction accuracy of an ML model is less relevant than batch recall statistics. A less trivial alignment challenge is the ubiquitous requirement across drug discovery for optimisation on multiple objectives – efficacy, safety, novelty, etc – and the need for AI systems to be evaluated across that spectrum of goals rather than on the often narrow selection of proxy optimisation objectives used in the training of the model.

Transparency

All the principles mentioned above are underpinned by the need for Transparency. It is the traditional viewpoint that the black-box nature of many ML models, such as neural networks, is an unavoidable price to pay for their effectiveness. However, in AI for science and, specifically, drug discovery, where AI and human experts work in concert towards the goal of a deeper understanding of nature, the need for an Explainable AI is non-negotiable. At BenevolentAI, we believe that ***good explanations foster better predictions***.

A classic example is the case of predicting therapeutic gene targets for a given disease. An AI model that provides evidence for each prediction, either in terms of gene regulatory pathways or selected explanatory text from the entire corpus of scientific literature, opens up further opportunities for scientists to propose further validation experiments or expand the scope of investigation based on criteria not captured by the AI system.

To this end, all things equal, we prioritise AI systems that explain their predictions. To lift the hood on the inner workings of our models, we focus on getting answers to the following questions:

- What feature perturbations are the model predictions most sensitive to?
- What information is the model actually using?
- Can we find similar examples to the predictions of the model?

In the case of neural networks, for instance, we approach these three challenges by performing gradient-based attribution, employing neural attention architectures, and k-nearest neighbours respectively. In addition to this, we train separate external interpretable linear models on neural network outputs, or on outputs under specific perturbations.