Ethics in Artificial Intelligence

Ethics as the Foundation of AI Risk Management

Artificial intelligence (AI) is poised to advance quality of life across the globe, bringing new benefits to people, organizations, and society. At Neo4j, we see the enormous potential of AI to enhance human life when used responsibly.

Through the National Institute of Standards and Technology (NIST), the U.S. government requested feedback on its first draft Artificial Intelligence Risk Management Framework. Neo4j submitted comments on the need to focus AI risk management on the prevention of human harm. Since ethics is the domain focused on minimizing human harm, we advised that **ethical principles should form the foundation of AI risk management**.

Considering ethics at the design stage makes sense because AI trained on data without ethical guardrails can cause unintended harm. AI systems reflect and even amplify the biases of their datasets, which can come from social disparities as well as data collection and labeling practices.

How Do You Embed Ethics in AI?

Governments and international organizations have been developing ethical codes for AI in recent years. Ethical codes for AI often draw from principles present in other legal frameworks, such as privacy and fairness.

Just as U.S. law is based on widely-accepted ethical principles, AI should be built with adherence to these principles to protect our civil liberties. NIST initially proposed **fairness, accountability, and transparency** as the principles that should guide AI risk management.

To put these or other principles into practice, we need to reflect on the purpose of the AI technology and what could go wrong during implementation. What do we know about the training data and how it was collected? Does the data reflect social disparities that could cause the system to draw erroneous conclusions? Is the training data representative of the populations that the AI technology will affect?

Once we've reached an understanding of the input data and implementation context, we can start thinking about what it would look like to put these principles into action. **Every AI use case is different, so it's expected that the way we embed ethics in each system will vary as well.**

An Example From Linkedin

Some years back, LinkedIn discovered that the algorithm it was using to match job candidates with opportunities was biased. Men were more active on the platform, so the algorithm learned to use gender as an indicator of job fitness.

Building fairness into AI systems like this one isn't an easy task, but it starts with a shared understanding of the principle. Fairness means that every person has an equal chance at receiving quality service, regardless of demographic.

Since every AI use has a unique purpose, fairness must be well-defined for each new system. We need to examine the dataset to assess how well it captures the experience of groups that will be affected by the system. Assessing the inclusivity of the dataset may lead us to collect more or different data, or clean it in ways that improve representativeness.

But operationalizing the concept of fairness doesn't end with the data. Even the machine learning (ML) techniques we choose reflect our understanding of fairness for each use case. In the LinkedIn example, treating each individual the same way (procedural fairness) isn't appropriate due to the bias against women in the dataset. Equal representation of men and women (group fairness) is a better option given the context.

Using a group fairness definition, we would allow for false positives over false negatives. A false positive means considering a woman who is not qualified; a false negative means rejecting a qualified woman candidate. Base rates in the data make it less likely that a woman would be selected in the first place, so we should minimize the rate of false negatives. Our familiarity with the dataset enables us to identify and mitigate bias by aligning our ML practice with the most appropriate use of the fairness principle.

Context-Driven, Ethical AI

Any AI risk management strategy must be based on a shared understanding of the ethical principles required for the use case. AI systems built without explicit reference to ethics will reproduce the biases of their datasets, creating risk for the individuals affected as well as the businesses that use them. By considering the ethical principles suited to each specific use case, businesses can begin to build safe, ethical AI systems.

For a deeper dive into this topic, read the full Neo4j response to the National Institute of Standards and Technology (NIST) on how ethics should inform AI design.

Questions about Neo4j? Contact us around the globe:

info@neo4j.com
neo4j.com/contact-us