Credit: Julie Su

Ethical AI: An Oxymoron or the Next Big Thing?

by Caleb A. Braun | BComm (BANA) '22

Artificial intelligence (AI) has developed rapidly in the past two decades. Accomplishments include the advent of autonomous driving (Kocić et al., 2019) and facial recognition (Parmar & Mehta, 2014), both of which are implementations of predictive models. A predictive model is a technique within AI that generates predictions by learning a relationship between input data and a target outcome measure (Menard, 2010). These models are increasingly used to influence decision-making in a variety of business applications (Bradlow et al., 2017; Collins et al., 2015; Raub, 2018) with substantial improvement over human judgment (Beam & Kohane, 2019). However, these accomplishments are accompanied by concerns over the unintended consequences when the decisions made based on the output of a model materially impact the well-being of individuals. This issue is increasingly recognized by technology-focused organizations, with most having developed or endorsed a set of ethical principles for AI (Fjeld et al., 2020). Only a minority of AI-focused organizations, however, "recognize the many risks of AI use, and fewer are working to reduce the risks" (McKinsey & Co., 2020, p. 9). When creating predictive models that materially impact the well-being of individuals, model creators should consider three factors: the bias in the model's training data, the transparency of the model, and how the model's performance will be validated.

Predictive models unknowingly built on inherently biased data can result in predictions that are unfavourable for minority subsets within the data, with embarrassing and dangerous outcomes (Chouldechova, 2017; Dressel & Farid, 2018). One example is COMPAS, a model used over the past two decades to predict the likelihood of recidivism of over one million American offenders (Angwin et al., 2016). Buolamwini and Gebru (2018) discussed bias in facial recognition algorithms, finding that three popular commercial facial recognition programs performed worse on women and those with dark-coloured skin. These biases can be difficult to detect (Edwards & Veale, 2018), and even if detected, are not easily eliminated. Proposed methods to reduce bias (Calders et al., 2009; Lum & Johndrow, 2016) are only effective if creators identify a specific bias; thus, unidentified biases can still linger. Regardless of whether the bias is identified prior to model building, any bias present in a model's training data will be transferred to the model (Lum & Johndrow, 2016) if no action is taken to mitigate it. Identifying and mitigating potential biases are crucial when developing predictive models, but additional measures are needed to develop ethical models.

A model's transparency is defined by the extent to which the prediction is explained by the model (Eddy et al., 2012), and it is inversely related to the model's complexity. Logistic regression is an example of a highly transparent model, as it shows the direction and magnitude of the relationship between each predictor variable and the response variable (Hastie et al., 2009). Transparent, logistic regression cannot capture complex, non-linear relationships, and thus may have lower levels of performance when compared to other models. In contrast to logistic regression, neural networks are a class of models that are designed to capture complex relationships, and thus achieve higher performance in some circumstances. Nevertheless, neural networks are not transparent (Hastie et al., 2009) as they lack a mechanism to explain how predictions are made. In contexts where predictions affect the well-being of people, predictive model creators should be cautious about using non-transparent models, notwithstanding their potentially superior performance. The European Union's General Data Protection Regulation was developed to protect individuals and their data (General Data Protection Regulation, 2016). This extends to the use of predictive models (Goodman & Flaxman, 2017). Article 13 discusses that the subject has the right to "meaningful information about the logic involved" in models used that "significantly affect (individuals)"(General Data Protection Regulation, 2016, p. 21). This is generally interpreted as the ability to give an individual an explanation of how specific aspects of their data affects their result (Edwards&Veale,2018). Although such regulations do not yet exist for private organizations within Canada, they still demonstrate the necessity for professionals to seriously consider the use of non-transparent models in human-affecting applications. The most conclusive understanding of how a model will perform upon deployment comes from validating performance on new data. Shneiderman discusses how "designers [need] to consider extreme situations and possible failures" (2020,p.6) and validate the model's performance on a wide range of new, independent data sets. Common performance measures include mean squared error (MSE) for a quantitative response variable (Sheiner & Beal, 1981) or sensitivity and specificity for a qualitative response (Altman & Bland, 1994).

No single required level of performance exists: the environment that the model will be used in determines the required performance level. Beyond statistical measures of performance, new data should be used to examine the societal consequences of the implementation of a predictive model (Corbett-Davies & Goel, 2018). In the example of predicting recidivism, model creators should consider how the model will be used to influence decision-making, and the subsequent impact on public safety and on those incarcerated. Using new data to validate a predictive model allows creators to anticipate both the model's real-world performance and potential consequences of implementation.

AI and predictive models are increasingly influencing decision-making and are crucial in the growth of large technology companies including Facebook (Hazelwood et al., 2018) and Google (Pichai, 2018). When developing predictive models that affect the well-being of individuals, creators have additional considerations to make to ensure the successful deployment of their model. To ensure equitable performance on sensitive subsets of individuals, the data used to create the predictive model should be evaluated for any inherent biases that may have negative impacts on those subsets. Model creators should consider the transparency requirements of the context in which their model will be used, and whether the benefits of a transparent model outweigh its potentially inferior performance. Finally, model creators must thoroughly evaluate their model's performance on new data, while examining the societal consequences of implementing the model. As regulations and ethical principles develop across international organizations, these considerations may even become requirements. Altogether, those predictive model creators who pre-emptively anticipate the consequences of implementing predictive models will develop models that are more easily implemented into the decision-making process, while being well-prepared for the future regulatory environment of AI.