Towards Socially Responsible AI: Cognitive Bias-Aware Multi-Objective Learning

December 11, 2021

```
\frame{\frametitle{figure 0}\begin{figure}[t]
\centering
\includegraphics[width=.75\columnwidth]{ceo.png}
\includegraphics[width=.24\columnwidth]{qc.png}
\caption{Examples of social bias in existing AI systems: 'Google
Image Search' under-representing women as CEOs (left), and
'Google query completion' prioritizing \emph{appearance} over
\emph{filmography} for a popular female actor (right).}
\label{fig:bias-in-existing-AI-tools}
\end{figure}}\frame{\frametitle{figure 1}\begin{figure}[t]
\centering
\includegraphics[width=.8\columnwidth]{bias-schematic.pdf}
\caption{Schematic of cognitive bias removal.}
\label{fig:schematic-bias}
\end{figure}}\frame{\frametitle{figure 2}\begin{figure}[t]
\centering
\includegraphics[width=.85\columnwidth]{debias-
architecture.pdf}
\caption{Schematic diagram of a neural network architecture for
```

\emph{jointly} learning the primary task objective (for effective prediction) and a set of debiasing tasks (for reducing the cognitive bias in these predictions).}
\label{fig:deBiasArch}
\end{figure}}\frame{\frametitle{figure 3}\begin{figure}[t]
\centering
\includegraphics[width=0.49\columnwidth]{biased_boundary.png}
\includegraphics[width=0.49\columnwidth]{debiased_boundary.png}
\caption{Illustrative example to visualize bias reduction with multi-objective learning (Equation \ref{eq:jointloss}). Predictions with a bias-agnostic classifier (logistic regression with L_2 regularization) are effective but exhibits a 'bias' in associating the green points with the top half of the plot area (left); whereas the multi-objective learning is able to reduce such a bias (right).
\label{fig:2d-data}}
\end{figure}}