Online Appendix to Algorithmic Social Engineering
Bo Cowgill and Megan T. Stevenson

Our setup assumed that the algorithm's predictions are more accurate than the judge's priors, in a way that is balanced across the different groups. We discuss the relaxing of this and other assumptions here. Since reliable data on actual crime doesn't exist, for instance, algorithms must use proxies such as arrest or conviction.

If the probability of arrest (conditional on criminal activity) differs across groups, then the algorithm's predictions would exhibit bias. If the developer manipulates the algorithm to increase accuracy across groups (de-biasing the predictions), this could be useful to decision-makers regardless of their preferences.

Several other extensions could also change the implications of our model. We mention a few below. We don't think these assumptions apply to judicial decision-making -- or many of the other settings for algorithmic social engineering. However, they may apply in certain circumstances. First, if algorithms themselves change the preferences of the decision-maker, then algorithmic social engineering might be more effective. We are not aware of research that measures whether algorithmic advice changes judges' preferences (rather than their information sets). However, career judges in an appointed or elected role are likely to have well-developed preferences.

Note that we are referring to a scenario in which the algorithm \emph{itself} changes preferences, not a scenario in which a jurisdiction implements penalties for failing to follow recommendations associated with the algorithm. This latter type of invention is an example of a policy that directly changes the decision-makers incentives, not their information set.

Second, if decision-makers are unaware that they are being manipulated (due to limited attention or the cost of information), then the strategic communication issues described above do not arise. However, decision-makers have many ways to observe the manipulation attempts, for example, in the public statements by advocates of new algorithms. Judges do not need to be subtle consumers of statistical data in order to detect manipulation attempts; they can simply read public statements.