Assessing Group Fairness with Social Welfare Optimization

John Hooker and **Derek Leben**
Carnegie Mellon University

Statistical parity metrics have been widely studied and endorsed in the AI community as a means of achieving fairness, but they suffer from at least two weaknesses. They disregard the actual welfare consequences of decisions and may therefore fail to achieve the kind of fairness that is desired for disadvantaged groups. In addition, they are often incompatible with each other, and there is no convincing justification for selecting one rather than another. This paper explores whether a broader conception of social justice, based on optimizing a social welfare function (SWF), can be useful for assessing various definitions of parity. We focus on the well-known alpha fairness SWF, which has been defended by axiomatic and bargaining arguments over a period of 70 years. We analyze the optimal solution and show that it can justify demographic parity or equalized odds under certain conditions, but frequently requires a departure from these types of parity. In addition, we find that predictive rate parity is of limited usefulness. These results suggest that optimization theory can shed light on the intensely discussed question of how to achieve group fairness in AI.