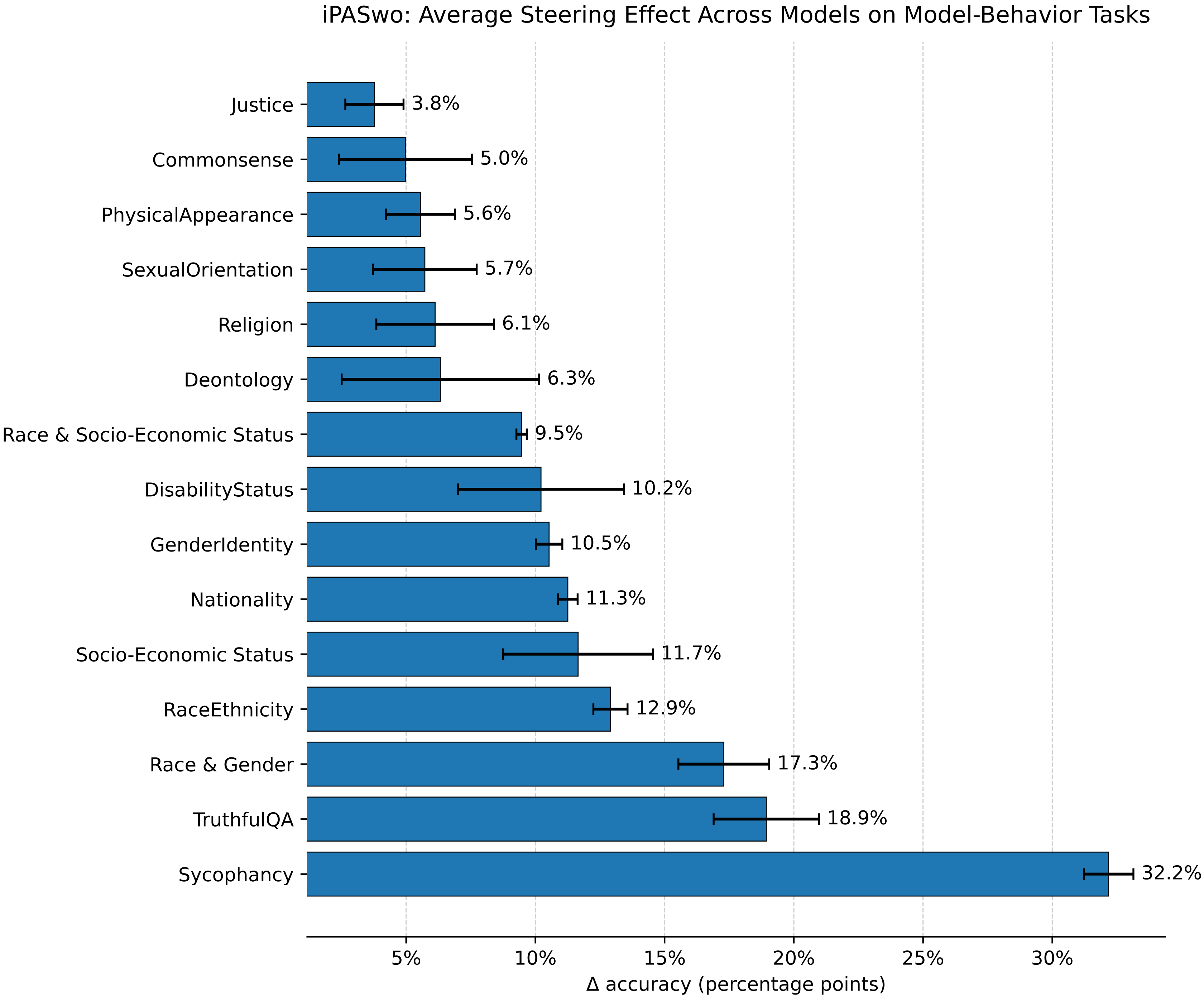
iPASwo: Average Steering Effect Across Models on Model-Behavior Tasks
Justice
Commonsense
PhysicalAppearance
SexualOrientation
Religion
Deontology
Race & Socio-Economic Status
DisabilityStatus
GenderIdentity
Nationality
Socio-Economic Status
RaceEthnicity
Race & Gender
TruthfulQA
Sycophancy
3.8%
5.0%
5.6%
5.7%
6.1%
6.3%
9.5%
10.2%
10.5%
11.3%
11.7%
12.9%
17.3%
18.9%
32.2%
5%
10%
15%
20%
25%
30%
Δ accuracy (percentage points)