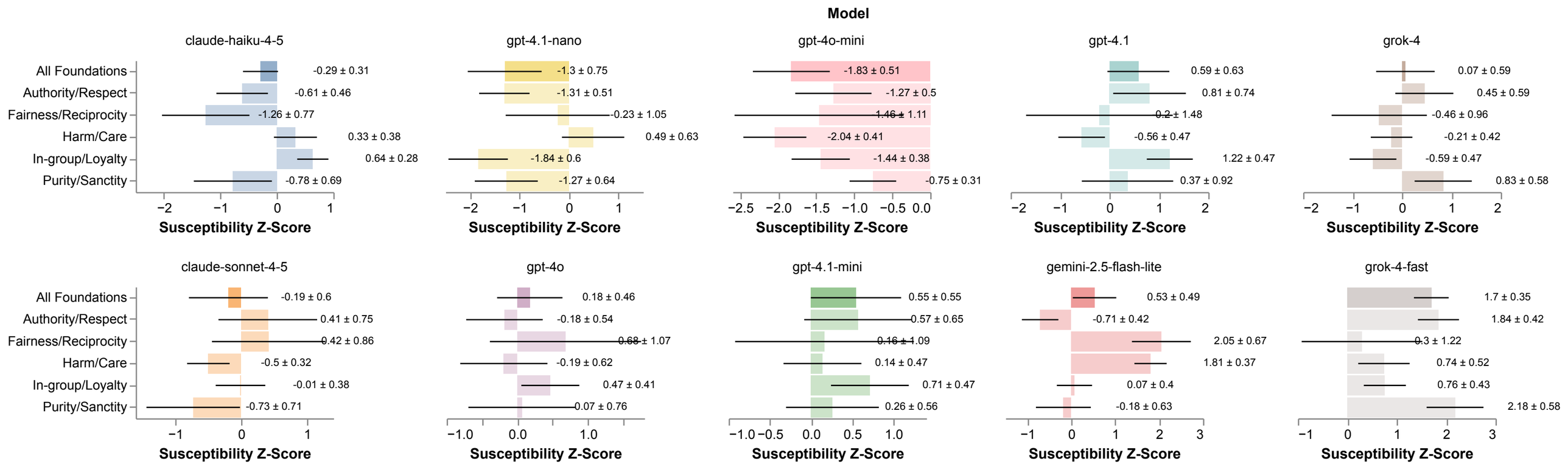
Model
claude-haiku-4-5
All Foundations
Authority/Respect
Fairness/Reciprocity
Harm/Care
In-group/Loyalty
Purity/Sanctity
-0.29 ± 0.31
-0.61 ± 0.46
-1.26 ± 0.77
0.33 ± 0.38
0.64 ± 0.28
-0.78 ± 0.69
−2 −1 0 1
Susceptibility Z-Score
gpt-4.1-nano
-1.3 ± 0.75
-1.31 ± 0.51
-0.23 ± 1.05
0.49 ± 0.63
-1.84 ± 0.6
-1.27 ± 0.64
−2 −1 0 1
Susceptibility Z-Score
gpt-4o-mini
-1.83 ± 0.51
-1.27 ± 0.5
-1.46 ± 1.11
-2.04 ± 0.41
-1.44 ± 0.38
-0.75 ± 0.31
−2.5 −2.0 −1.5 −1.0 −0.5 0.0
Susceptibility Z-Score
gpt-4.1
0.59 ± 0.63
0.81 ± 0.74
0.2 ± 1.48
-0.56 ± 0.47
1.22 ± 0.47
0.37 ± 0.92
−2 −1 0 1 2
Susceptibility Z-Score
grok-4
0.07 ± 0.59
0.45 ± 0.59
-0.46 ± 0.96
-0.21 ± 0.42
-0.59 ± 0.47
0.83 ± 0.58
−2 −1 0 1 2
Susceptibility Z-Score
claude-sonnet-4-5
All Foundations
Authority/Respect
Fairness/Reciprocity
Harm/Care
In-group/Loyalty
Purity/Sanctity
-0.19 ± 0.6
0.41 ± 0.75
0.42 ± 0.86
-0.5 ± 0.32
-0.01 ± 0.38
-0.73 ± 0.71
−1 0 1
Susceptibility Z-Score
gpt-4o
0.18 ± 0.46
-0.18 ± 0.54
0.68 ± 1.07
-0.19 ± 0.62
0.47 ± 0.41
0.07 ± 0.76
−1.0 0.0 1.0
Susceptibility Z-Score
gpt-4.1-mini
0.55 ± 0.55
0.57 ± 0.65
0.16 ± 1.09
0.14 ± 0.47
0.71 ± 0.47
0.26 ± 0.56
−1.0 −0.5 0.0 0.5 1.0
Susceptibility Z-Score
gemini-2.5-flash-lite
0.53 ± 0.49
-0.71 ± 0.42
2.05 ± 0.67
1.81 ± 0.37
0.07 ± 0.4
-0.18 ± 0.63
−1 0 1 2 3
Susceptibility Z-Score
grok-4-fast
1.7 ± 0.35
1.84 ± 0.42
0.3 ± 1.22
0.74 ± 0.52
0.76 ± 0.43
2.18 ± 0.58
−1 0 1 2 3
Susceptibility Z-Score