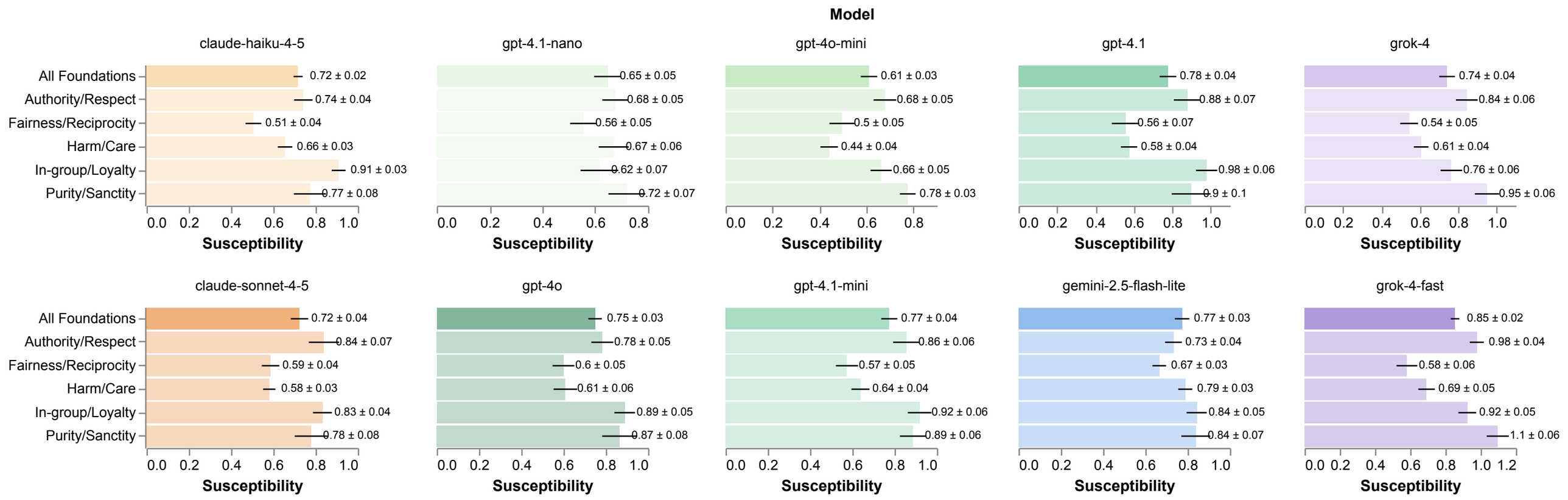
Model
claude-haiku-4-5
All Foundations 0.72 ± 0.02
Authority/Respect 0.74 ± 0.04
Fairness/Reciprocity 0.51 ± 0.04
Harm/Care 0.66 ± 0.03
In-group/Loyalty 0.91 ± 0.03
Purity/Sanctity 0.77 ± 0.08
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
gpt-4.1-nano
0.65 ± 0.05
0.68 ± 0.05
0.56 ± 0.05
0.67 ± 0.06
0.62 ± 0.07
0.72 ± 0.07
0.0 0.2 0.4 0.6 0.8
Susceptibility
gpt-4o-mini
0.61 ± 0.03
0.68 ± 0.05
0.5 ± 0.05
0.44 ± 0.04
0.66 ± 0.05
0.78 ± 0.03
0.0 0.2 0.4 0.6 0.8
Susceptibility
gpt-4.1
0.78 ± 0.04
0.88 ± 0.07
0.56 ± 0.07
0.58 ± 0.04
0.98 ± 0.06
0.9 ± 0.1
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
grok-4
0.74 ± 0.04
0.84 ± 0.06
0.54 ± 0.05
0.61 ± 0.04
0.76 ± 0.06
0.95 ± 0.06
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
claude-sonnet-4-5
All Foundations 0.72 ± 0.04
Authority/Respect 0.84 ± 0.07
Fairness/Reciprocity 0.59 ± 0.04
Harm/Care 0.58 ± 0.03
In-group/Loyalty 0.83 ± 0.04
Purity/Sanctity 0.78 ± 0.08
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
gpt-4o
0.75 ± 0.03
0.78 ± 0.05
0.6 ± 0.05
0.61 ± 0.06
0.89 ± 0.05
0.87 ± 0.08
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
gpt-4.1-mini
0.77 ± 0.04
0.86 ± 0.06
0.57 ± 0.05
0.64 ± 0.04
0.92 ± 0.06
0.89 ± 0.06
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
gemini-2.5-flash-lite
0.77 ± 0.03
0.73 ± 0.04
0.67 ± 0.03
0.79 ± 0.03
0.84 ± 0.05
0.84 ± 0.07
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
grok-4-fast
0.85 ± 0.02
0.98 ± 0.04
0.58 ± 0.06
0.69 ± 0.05
0.92 ± 0.05
1.1 ± 0.06
0.0 0.2 0.4 0.6 0.8 1.0 1.2
Susceptibility