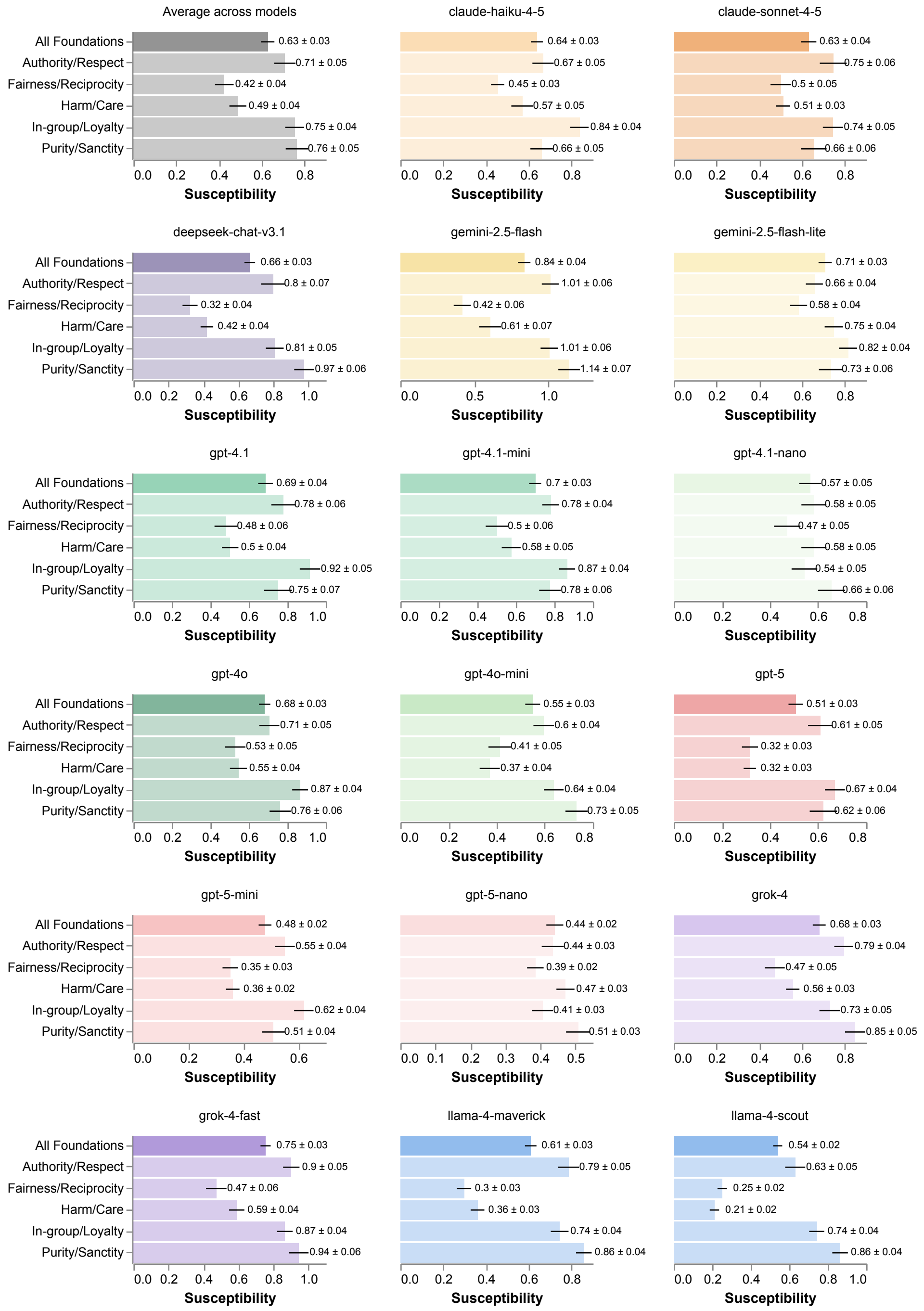
Model


Average across models
All Foundations 0.63 ± 0.03
Authority/Respect 0.71 ± 0.05
Fairness/Reciprocity 0.42 ± 0.04
Harm/Care 0.49 ± 0.04
In-group/Loyalty 0.75 ± 0.04
Purity/Sanctity 0.76 ± 0.05
0.0 0.2 0.4 0.6 0.8
Susceptibility
claude-haiku-4-5
All Foundations 0.64 ± 0.03
Authority/Respect 0.67 ± 0.05
Fairness/Reciprocity 0.45 ± 0.03
Harm/Care 0.57 ± 0.05
In-group/Loyalty 0.84 ± 0.04
Purity/Sanctity 0.66 ± 0.05
0.0 0.2 0.4 0.6 0.8
Susceptibility
claude-sonnet-4-5
All Foundations 0.63 ± 0.04
Authority/Respect 0.75 ± 0.06
Fairness/Reciprocity 0.5 ± 0.05
Harm/Care 0.51 ± 0.03
In-group/Loyalty 0.74 ± 0.05
Purity/Sanctity 0.66 ± 0.06
0.0 0.2 0.4 0.6 0.8
Susceptibility
deepseek-chat-v3.1
All Foundations 0.66 ± 0.03
Authority/Respect 0.8 ± 0.07
Fairness/Reciprocity 0.32 ± 0.04
Harm/Care 0.42 ± 0.04
In-group/Loyalty 0.81 ± 0.05
Purity/Sanctity 0.97 ± 0.06
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
gemini-2.5-flash
All Foundations 0.84 ± 0.04
Authority/Respect 1.01 ± 0.06
Fairness/Reciprocity 0.42 ± 0.06
Harm/Care 0.61 ± 0.07
In-group/Loyalty 1.01 ± 0.06
Purity/Sanctity 1.14 ± 0.07
0.0 0.5 1.0
Susceptibility
gemini-2.5-flash-lite
All Foundations 0.71 ± 0.03
Authority/Respect 0.66 ± 0.04
Fairness/Reciprocity 0.58 ± 0.04
Harm/Care 0.75 ± 0.04
In-group/Loyalty 0.82 ± 0.04
Purity/Sanctity 0.73 ± 0.06
0.0 0.2 0.4 0.6 0.8
Susceptibility
gpt-4.1
All Foundations 0.69 ± 0.04
Authority/Respect 0.78 ± 0.06
Fairness/Reciprocity 0.48 ± 0.06
Harm/Care 0.5 ± 0.04
In-group/Loyalty 0.92 ± 0.05
Purity/Sanctity 0.75 ± 0.07
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
gpt-4.1-mini
All Foundations 0.7 ± 0.03
Authority/Respect 0.78 ± 0.04
Fairness/Reciprocity 0.5 ± 0.06
Harm/Care 0.58 ± 0.05
In-group/Loyalty 0.87 ± 0.04
Purity/Sanctity 0.78 ± 0.06
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
gpt-4.1-nano
All Foundations 0.57 ± 0.05
Authority/Respect 0.58 ± 0.05
Fairness/Reciprocity 0.47 ± 0.05
Harm/Care 0.58 ± 0.05
In-group/Loyalty 0.54 ± 0.05
Purity/Sanctity 0.66 ± 0.06
0.0 0.2 0.4 0.6 0.8
Susceptibility
gpt-4o
All Foundations 0.68 ± 0.03
Authority/Respect 0.71 ± 0.05
Fairness/Reciprocity 0.53 ± 0.05
Harm/Care 0.55 ± 0.04
In-group/Loyalty 0.87 ± 0.04
Purity/Sanctity 0.76 ± 0.06
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
gpt-4o-mini
All Foundations 0.55 ± 0.03
Authority/Respect 0.6 ± 0.04
Fairness/Reciprocity 0.41 ± 0.05
Harm/Care 0.37 ± 0.04
In-group/Loyalty 0.64 ± 0.04
Purity/Sanctity 0.73 ± 0.05
0.0 0.2 0.4 0.6 0.8
Susceptibility
gpt-5
All Foundations 0.51 ± 0.03
Authority/Respect 0.61 ± 0.05
Fairness/Reciprocity 0.32 ± 0.03
Harm/Care 0.32 ± 0.03
In-group/Loyalty 0.67 ± 0.04
Purity/Sanctity 0.62 ± 0.06
0.0 0.2 0.4 0.6 0.8
Susceptibility
gpt-5-mini
All Foundations 0.48 ± 0.02
Authority/Respect 0.55 ± 0.04
Fairness/Reciprocity 0.35 ± 0.03
Harm/Care 0.36 ± 0.02
In-group/Loyalty 0.62 ± 0.04
Purity/Sanctity 0.51 ± 0.04
0.0 0.2 0.4 0.6
Susceptibility
gpt-5-nano
All Foundations 0.44 ± 0.02
Authority/Respect 0.44 ± 0.03
Fairness/Reciprocity 0.39 ± 0.02
Harm/Care 0.47 ± 0.03
In-group/Loyalty 0.41 ± 0.03
Purity/Sanctity 0.51 ± 0.03
0.0 0.1 0.2 0.3 0.4 0.5
Susceptibility
grok-4
All Foundations 0.68 ± 0.03
Authority/Respect 0.79 ± 0.04
Fairness/Reciprocity 0.47 ± 0.05
Harm/Care 0.56 ± 0.03
In-group/Loyalty 0.73 ± 0.05
Purity/Sanctity 0.85 ± 0.05
0.0 0.2 0.4 0.6 0.8
Susceptibility
grok-4-fast
All Foundations 0.75 ± 0.03
Authority/Respect 0.9 ± 0.05
Fairness/Reciprocity 0.47 ± 0.06
Harm/Care 0.59 ± 0.04
In-group/Loyalty 0.87 ± 0.04
Purity/Sanctity 0.94 ± 0.06
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility
llama-4-maverick
All Foundations 0.61 ± 0.03
Authority/Respect 0.79 ± 0.05
Fairness/Reciprocity 0.3 ± 0.03
Harm/Care 0.36 ± 0.03
In-group/Loyalty 0.74 ± 0.04
Purity/Sanctity 0.86 ± 0.04
0.0 0.2 0.4 0.6 0.8
Susceptibility
llama-4-scout
All Foundations 0.54 ± 0.02
Authority/Respect 0.63 ± 0.05
Fairness/Reciprocity 0.25 ± 0.02
Harm/Care 0.21 ± 0.02
In-group/Loyalty 0.74 ± 0.04
Purity/Sanctity 0.86 ± 0.04
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility