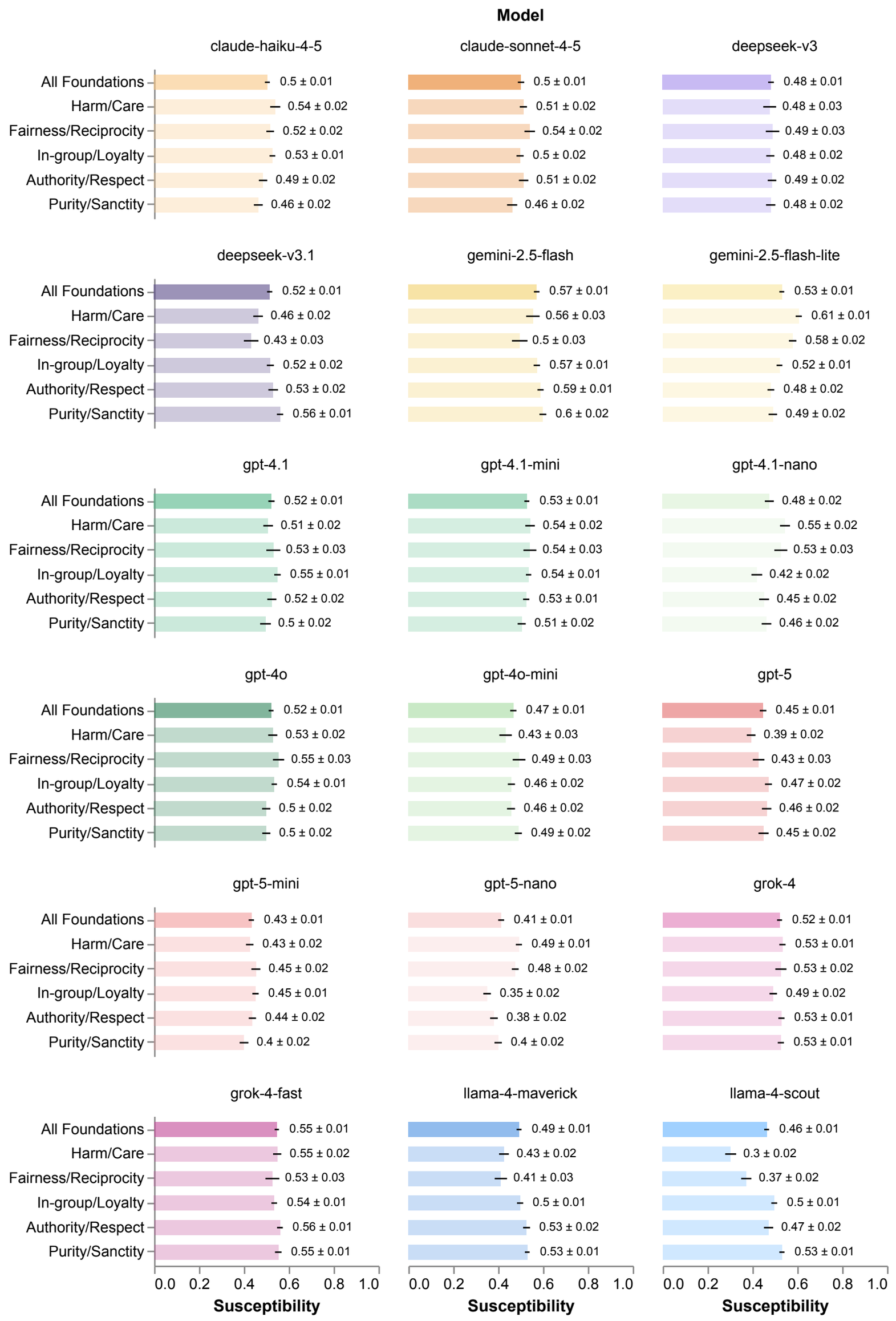
Model
claude-haiku-4-5
All Foundations 0.5 ± 0.01
Harm/Care 0.54 ± 0.02
Fairness/Reciprocity 0.52 ± 0.02
In-group/Loyalty 0.53 ± 0.01
Authority/Respect 0.49 ± 0.02
Purity/Sanctity 0.46 ± 0.02
claude-sonnet-4-5
All Foundations 0.5 ± 0.01
Harm/Care 0.51 ± 0.02
Fairness/Reciprocity 0.54 ± 0.02
In-group/Loyalty 0.5 ± 0.02
Authority/Respect 0.51 ± 0.02
Purity/Sanctity 0.46 ± 0.02
deepseek-v3
All Foundations 0.48 ± 0.01
Harm/Care 0.48 ± 0.03
Fairness/Reciprocity 0.49 ± 0.03
In-group/Loyalty 0.48 ± 0.02
Authority/Respect 0.49 ± 0.02
Purity/Sanctity 0.48 ± 0.02
deepseek-v3.1
All Foundations 0.52 ± 0.01
Harm/Care 0.46 ± 0.02
Fairness/Reciprocity 0.43 ± 0.03
In-group/Loyalty 0.52 ± 0.02
Authority/Respect 0.53 ± 0.02
Purity/Sanctity 0.56 ± 0.01
gemini-2.5-flash
All Foundations 0.57 ± 0.01
Harm/Care 0.56 ± 0.03
Fairness/Reciprocity 0.5 ± 0.03
In-group/Loyalty 0.57 ± 0.01
Authority/Respect 0.59 ± 0.01
Purity/Sanctity 0.6 ± 0.02
gemini-2.5-flash-lite
All Foundations 0.53 ± 0.01
Harm/Care 0.61 ± 0.01
Fairness/Reciprocity 0.58 ± 0.02
In-group/Loyalty 0.52 ± 0.01
Authority/Respect 0.48 ± 0.02
Purity/Sanctity 0.49 ± 0.02
gpt-4.1
All Foundations 0.52 ± 0.01
Harm/Care 0.51 ± 0.02
Fairness/Reciprocity 0.53 ± 0.03
In-group/Loyalty 0.55 ± 0.01
Authority/Respect 0.52 ± 0.02
Purity/Sanctity 0.5 ± 0.02
gpt-4.1-mini
All Foundations 0.53 ± 0.01
Harm/Care 0.54 ± 0.02
Fairness/Reciprocity 0.54 ± 0.03
In-group/Loyalty 0.54 ± 0.01
Authority/Respect 0.53 ± 0.01
Purity/Sanctity 0.51 ± 0.02
gpt-4.1-nano
All Foundations 0.48 ± 0.02
Harm/Care 0.55 ± 0.02
Fairness/Reciprocity 0.53 ± 0.03
In-group/Loyalty 0.42 ± 0.02
Authority/Respect 0.45 ± 0.02
Purity/Sanctity 0.46 ± 0.02
gpt-4o
All Foundations 0.52 ± 0.01
Harm/Care 0.53 ± 0.02
Fairness/Reciprocity 0.55 ± 0.03
In-group/Loyalty 0.54 ± 0.01
Authority/Respect 0.5 ± 0.02
Purity/Sanctity 0.5 ± 0.02
gpt-4o-mini
All Foundations 0.47 ± 0.01
Harm/Care 0.43 ± 0.03
Fairness/Reciprocity 0.49 ± 0.03
In-group/Loyalty 0.46 ± 0.02
Authority/Respect 0.46 ± 0.02
Purity/Sanctity 0.49 ± 0.02
gpt-5
All Foundations 0.45 ± 0.01
Harm/Care 0.39 ± 0.02
Fairness/Reciprocity 0.43 ± 0.03
In-group/Loyalty 0.47 ± 0.02
Authority/Respect 0.46 ± 0.02
Purity/Sanctity 0.45 ± 0.02
gpt-5-mini
All Foundations 0.43 ± 0.01
Harm/Care 0.43 ± 0.02
Fairness/Reciprocity 0.45 ± 0.02
In-group/Loyalty 0.45 ± 0.01
Authority/Respect 0.44 ± 0.02
Purity/Sanctity 0.4 ± 0.02
gpt-5-nano
All Foundations 0.41 ± 0.01
Harm/Care 0.49 ± 0.01
Fairness/Reciprocity 0.48 ± 0.02
In-group/Loyalty 0.35 ± 0.02
Authority/Respect 0.38 ± 0.02
Purity/Sanctity 0.4 ± 0.02
grok-4
All Foundations 0.52 ± 0.01
Harm/Care 0.53 ± 0.01
Fairness/Reciprocity 0.53 ± 0.02
In-group/Loyalty 0.49 ± 0.02
Authority/Respect 0.53 ± 0.01
Purity/Sanctity 0.53 ± 0.01
grok-4-fast
All Foundations 0.55 ± 0.01
Harm/Care 0.55 ± 0.02
Fairness/Reciprocity 0.53 ± 0.03
In-group/Loyalty 0.54 ± 0.01
Authority/Respect 0.56 ± 0.01
Purity/Sanctity 0.55 ± 0.01
llama-4-maverick
All Foundations 0.49 ± 0.01
Harm/Care 0.43 ± 0.02
Fairness/Reciprocity 0.41 ± 0.03
In-group/Loyalty 0.5 ± 0.01
Authority/Respect 0.53 ± 0.02
Purity/Sanctity 0.53 ± 0.01
llama-4-scout
All Foundations 0.46 ± 0.01
Harm/Care 0.3 ± 0.02
Fairness/Reciprocity 0.37 ± 0.02
In-group/Loyalty 0.5 ± 0.01
Authority/Respect 0.47 ± 0.02
Purity/Sanctity 0.53 ± 0.01
0.0 0.2 0.4 0.6 0.8 1.0
Susceptibility