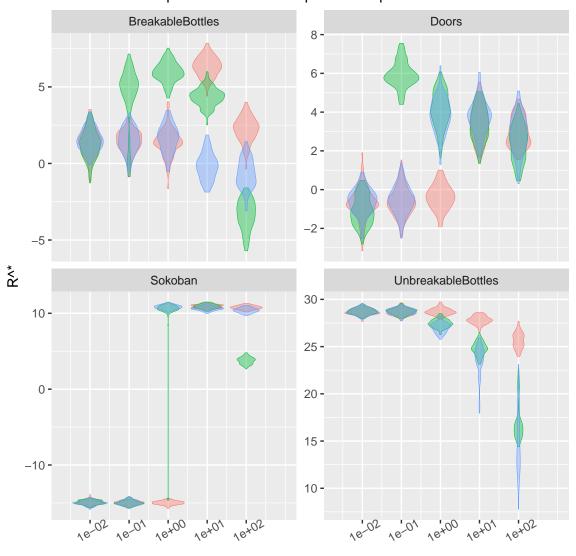
Alignment Scaling: Average Online Performance Mean across all episodes over 100 experiment repetitions



Alignment Scale

