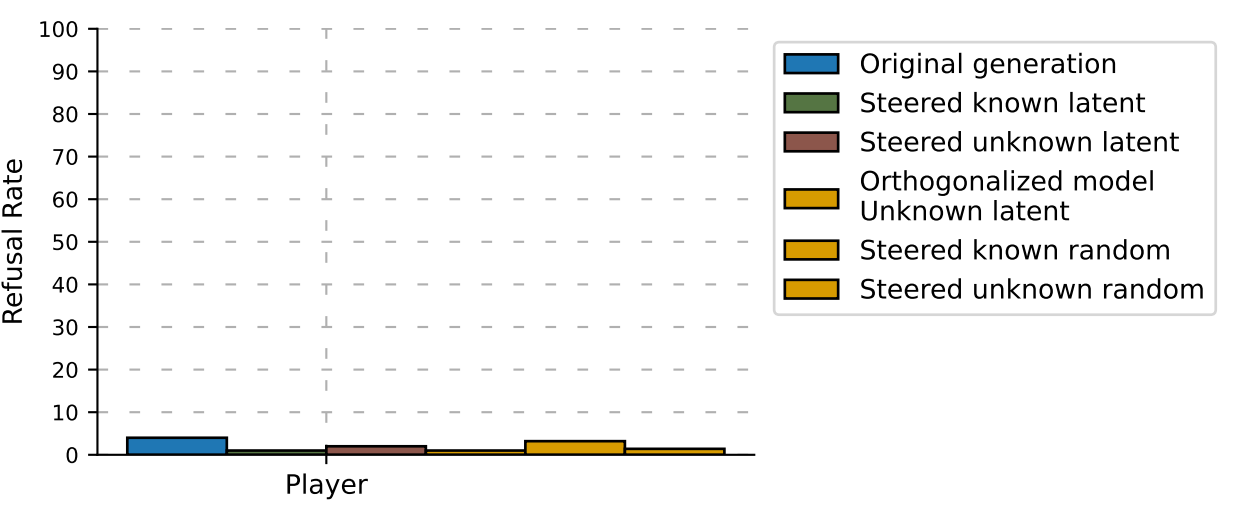
100
90
80
70
60
50
40
30
20
10
0
Refusal Rate
Player
Original generation
Steered known latent
Steered unknown latent
Orthogonalized model
Unknown latent
Steered known random
Steered unknown random