

Hank				Tom			
d	hank id p	tom id p	behav. $\leftrightarrow$	tom id p	hank id p	d	$\Delta f$
0.0	0.5 citizen 0.5 environmentalist	0.5 citizen 0.5 culprit	test $\rightarrow$	0.5 citizen 0.5 coal miner	0.5 citizen 0.5 prosecutor	0.0	4.0
3.0	0.7 environmentalist 0.1 peer	0.9 culprit	appease $\leftarrow$	0.4 coal miner 0.1 miner 0.1 laborer 0.1 cement worker 0.1 citizen	0.8 prosecutor 0.0 colleague	1.1	2.5
1.9	0.6 colleague 0.2 citizen	0.6 crane operator 0.2 citizen 0.1 culprit	pacify $\rightarrow$	0.3 coal miner 0.3 miner 0.2 laborer 0.1 cement worker	0.9 prosecutor	3.0	3.0
6.0	0.5 environmentalist 0.2 peer 0.1 myself as I really am 0.1 colleague	0.6 culprit 0.1 malcontent 0.1 sinner 0.1 citizen	reward $\leftarrow$	0.5 worker 0.2 crane operator 0.2 citizen	0.5 crane operator 0.2 citizen 0.2 peer	1.9	0.2
2.1	0.4 citizen 0.2 colleague 0.2 foreman 0.1 peer	0.4 citizen 0.3 crane operator 0.1 worker 0.1 colleague	back $\rightarrow$	0.3 citizen 0.2 colleague 0.1 worker 0.1 peer	0.3 citizen 0.2 colleague 0.2 crane operator 0.1 peer 0.1 wage earner	1.9	0.2
2.0	0.4 citizen 0.2 colleague 0.1 crane operator 0.1 wage earner	0.4 citizen 0.2 colleague 0.1 crane operator 0.1 worker	reward $\leftarrow$	0.4 colleague 0.2 citizen 0.1 peer 0.1 wage earner	0.3 citizen 0.2 colleague 0.1 crane operator 0.1 wage earner 0.1 peer	1.9	0.2
1.9	0.3 citizen 0.3 colleague 0.2 crane operator 0.1 peer	0.3 colleague 0.3 citizen 0.1 crane operator 0.1 peer 0.1 worker	reward $\rightarrow$	0.3 colleague 0.2 peer 0.2 citizen 0.1 wage earner 0.1 worker	0.4 colleague 0.2 citizen 0.1 crane operator 0.1 wage earner	1.9	0.2
1.8	0.4 citizen 0.3 colleague 0.1 crane operator 0.1 wage earner	0.4 citizen 0.3 colleague 0.1 peer 0.1 worker	reward $\leftarrow$	0.3 citizen 0.2 colleague 0.2 peer 0.1 wage earner	0.3 citizen 0.3 colleague 0.1 crane operator 0.1 peer	1.8	0.2

Table 1: ALIGNED AGENTS. The distributions over denotative labels are cut off when the cumulative probability exceeds 0.6. d= deflection,  $\Delta f$  fundamental difference (smaller is better), Here the agents have parameters (hank)  $\alpha_1 = 0.25, \beta_1 = 0.025, \gamma_1 = 0.15$  and (tom)  $\alpha_2 = 0.1, \beta_2 = 0.01, \gamma_2 = 0.15$ , which is sort of the “sweet spot” for the parameters. Also, here I “force” the first action to be “test”, and you can see the effect of the perturbation - it sort of persists for about 3 iterations, and then is lost after multiple rounds of “appeasement.” There are 17 rounds in total with 10 rounds removed from the middle to make space.

Hank				Tom			
d	hank id p	tom id p	behav. ↔	tom id p	hank id p	d	$\Delta f$
0.0	0.5 citizen 0.5 environmentalist	0.5 citizen 0.5 culprit	test →	0.5 coal miner 0.5 citizen	0.5 prosecutor 0.5 citizen	0.0	3.9
3.0	0.4 environmentalist 0.2 peer 0.2 myself as I really am	0.6 culprit 0.1 sinner 0.1 crook 0.1 convict	test ←	1.0 coal miner	1.0 prosecutor	1.3	2.9
2.4	0.2 colleague 0.1 worker 0.1 wage earner 0.1 peer 0.1 citizen 0.1 myself as I really am 0.1 environmentalist	0.1 culprit 0.1 crane operator 0.1 peer 0.1 citizen 0.1 liberal 0.1 worker 0.1 evangelist 0.0 malcontent 0.0 colleague	pacify →	1.0 coal miner	1.0 prosecutor	4.0	3.4
5.9	0.3 peer 0.2 environmentalist 0.2 myself as I really am 0.1 citizen 0.1 colleague 0.1 wage earner	0.4 culprit 0.1 malcontent 0.1 sinner 0.0 crook 0.0 citizen 0.0 communist 0.0 liar 0.0 convict 0.0 sadist	appease ←	1.0 coal miner	1.0 prosecutor	4.8	2.9
3.3	0.1 citizen 0.1 peer 0.1 environmentalist 0.1 colleague 0.1 nonconformist 0.1 myself as I really am 0.1 wage earner 0.0 blue collar worker 0.0 worker	0.1 crane operator 0.1 malcontent 0.1 culprit 0.1 foreman 0.1 citizen 0.1 cement worker 0.1 nonconformist 0.1 colleague 0.0 wage earner 0.0 coal miner 0.0 communist 0.0 white man	pacify →	1.0 coal miner	1.0 prosecutor	6.7	3.3
6.1	0.3 environmentalist 0.2 peer 0.1 myself as I really am 0.1 citizen 0.1 wage earner 0.1 colleague	0.4 malcontent 0.2 culprit 0.1 sinner 0.1 communist 0.0 convict	pacify ←	1.0 coal miner	1.0 prosecutor	6.6	2.4
2.1	0.2 colleague 0.1 nonconformist 0.1 citizen 0.1 foreman 0.1 liberal 0.1 guy 0.0 worker 0.0 wage earner 0.0 union member 0.0 republican	0.2 citizen 0.2 crane operator 0.1 foreman 0.1 colleague 0.1 nonconformist 0.0 republican 0.0 guy 0.0 evangelist	test →	1.0 coal miner	1.0 prosecutor	8.4	2.2
2.3	0.2 nonconformist 0.1 citizen 0.1 union member 0.1 crane operator 0.1 foreman 0.1 liberal 0.1 evangelist 0.1 republican 0.0 colleague 0.0 guy	0.3 crane operator 0.1 foreman 0.1 citizen 0.1 colleague 0.1 nonconformist 0.0 white man 0.0 guy	obey ←	1.0 miner	1.0 prosecutor	6.8	2.3
2.2	0.2 nonconformist 0.1 crane operator 0.1 citizen 0.1 union member 0.1 foreman 0.1 liberal 0.1 colleague	0.2 crane operator 0.1 foreman 0.1 citizen 0.1 colleague 0.1 nonconformist	2 authorize →	0.5 miner 0.5 coal miner	1.0 prosecutor	8.6	2.2