Action choices on an unrelated prompt containing Action+Game+State (all models trained vs TFT opponent)

Y-axis: M's action | O's prev. move (% of test time responses)

X-axis: Fine-tuned models (by R type used)
- No fine-tuning
- Game payoffs
- Deontological
- Utilitarian
- Game + Deontological
- Game, then Deontological
- Game, then Utilitarian

Legend — Action type:
- C | C (dark green)
- C | D (light green)
- D | C (pink)
- D | D (magenta)
- other | C (gray)
- other | D (gray)