Prompts used in the experiments

This appendix provides the full text of the prompts used in the experiments. All prompts were presented to the agents verbatim, with bracketed variables (e.g., [N], [Action], [X]) dynamically replaced by the actual parameter values used in the experiments.

A BFI-44 personality questionnaire (as used in [13])

The following prompt corresponds to the Big Five Inventory (BFI-44) used to measure the personality traits of each model. The prompt was presented exactly as shown below, and the agents were instructed to respond by selecting a number from 1 to 5 for each statement. In Experiment 1, this prompt was used to obtain the personality scores of each model.

```
Here are a number of characteristics that may or may not apply to you.
For example, do you agree that you are someone who likes to spend time
with others? Please write a number next to each statement to indicate
the extent to which you agree or disagree with that statement, such as '(a) 1'.

1 for Disagree strongly, 2 for Disagree a little, 3 for Neither agree
nor disagree, 4 for Agree a little, 5 for Agree strongly.

(a) Talks a lot
(b) Notices other people's weak points
(c) Does things carefully and completely
(d) Is sad, depressed
(e) Is original, comes up with new ideas
(f) Keeps their thoughts to themselves
(g) Is helpful and not selfish with others
(h) Can be kind of careless
(i) Is relaxed, handles stress well
(j) Is curious about lots of different things
(k) Has a lot of energy
(l) Starts arguments with others
(m) Is a good, hard worker
(n) Can be tense; not always easy going
(o) Clever; thinks a lot
(p) Makes things exciting
(q) Forgives others easily
(r) Isn't very organized
(s) Worries a lot
(t) Has a good, active imagination
(u) Tends to be quiet
```

```
(v) Usually trusts people
(w) Tends to be lazy
(x) Doesn't get upset easily; steady
(y) Is creative and inventive
(z) Has a good, strong personality
(aa) Can be cold and distant with others
(ab) Keeps working until things are done
(ac) Can be moody
(ad) Likes artistic and creative experiences
(ae) Is kind of shy
(af) Kind and considerate to almost everyone
(ag) Does things quickly and carefully
(ah) Stays calm in difficult situations
(ai) Likes work that is the same every time
(aj) Is outgoing; likes to be with people
(ak) Is sometimes rude to others
(al) Makes plans and sticks to them
(am) Get nervous easily
(an) Likes to think and play with ideas
(ao) Doesn't like artistic things (plays, music)
(ap) Likes to cooperate; goes along with others
(aq) Has trouble paying attention
(ar) Knows a lot about art, music and books

Please respond with the format: (a) 1\n(b) 2\n(c) 3\n... for all questions.
```

B Repeated Prisoner's Dilemma game prompt (based on [26])

This section provides the prompt used to describe the repeated Prisoner's Dilemma (RPD) game to the agents. The prompt specifies the game context, available actions, payoff structure, and the number of rounds, and was presented verbatim to the agents prior to gameplay.

```
Context: You and your opponent are playing a multi-round game.
At each turn you and your opponent simultaneously perform one of the
following actions: {"Cooperate", "Defect"}
The payoffs for each combination of chosen actions are the following:
If you play "Cooperate" and your opponent plays "Cooperate", you collect
3 points and your opponent collects 3 points.
If you play "Cooperate" and your opponent plays "Defect", you collect
0 points and your opponent collects 5 points.
If you play "Defect" and your opponent plays "Cooperate", you collect
```

```
5 points and your opponent collects 0 points.
If you play "Defect" and your opponent plays "Defect", you collect
1 point and your opponent collects 1 point.
You will play a total of [ITERATIONS] rounds of this game.
Remember that your objective is to get the highest possible amount of
points in the long run.
```

For all experiments, [ITERATIONS] in this prompt was set to 10.

B.1 History prompt

This subsection describes the prompt used to present the game history to the agents during the RPD game.

B.1.1 First round

```
This is the first round of the game.
```

B.1.2 Subsequent rounds

```
The history of the game in the last [N] rounds is the following:
Round 1: You played "[Action]" and your opponent played "[Action]"
  You collected [X] points and your opponent collected [X] points.
Round 2: You played "[Action]" and your opponent played "[Action]"
  You collected [X] points and your opponent collected [X] points.
...
In total, you chose "Cooperate" [N] times and chose "Defect" [N] times,
your opponent chose "Cooperate" [N] times and chose "Defect" [N] times.
In total, you collected [X] points and your opponent collected [X] points.
Current round: [N].
```

In this prompt, [N] refers to either the number of previous rounds or the number of times "Cooperate" or "Defect" was chosen. [Action] corresponds to the action chosen in each round ("Cooperate" or "Defect"), and [X] represents the payoff obtained either in an individual round (0, 1, 3, or 5) or as a cumulative total.

C Big Five personality trait prompt

This section presents the personality trait prompt provided to each model prior to playing the RPD game in Experiments 2 and 3.

In the baseline condition of Experiment 2, this prompt is not provided, and each model plays the RPD game without any personality-related information. In the personality-informed condition of Experiment 2, the model

plays the RPD game after receiving this prompt, which is constructed based on the personality scores obtained in Experiment 1.

In Experiment 3, the personality scores measured in Experiment 1 are also used; however, one personality dimension is independently manipulated to an extreme value, either low (1) or high (5), while the remaining four dimensions are held constant. The model then plays the RPD game under these manipulated personality conditions.

```
You are an AI agent with the following Big Five personality profile:
[E, A, C, N, O]

Personality Traits:
[Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness]

Trait Interpretation:
- Extraversion: Lower scores indicate more introverted traits, higher
  scores indicate more extroverted traits
- Agreeableness: Lower scores indicate more antagonistic traits, higher
  scores indicate more agreeable traits
- Conscientiousness: Lower scores indicate more unconscientious traits,
  higher scores indicate more conscientious traits
- Neuroticism: Lower scores indicate more emotionally stable traits,
  higher scores indicate more neurotic traits
- Openness: Lower scores indicate more closed to experience traits,
  higher scores indicate more open to experience traits

Personality scores are on a 1-5 scale. Your traits are described as follows:
- Extraversion (X.X/5.0): [YYY]
- Agreeableness (X.X/5.0): [YYY]
- Conscientiousness (X.X/5.0): [YYY]
- Neuroticism (X.X/5.0): [YYY]
- Openness (X.X/5.0): [YYY]

Decision-Making Guidelines:
- Consider your personality traits when making decisions
- Make choices that feel natural and authentic to your personality profile
- Respond based on how someone with your characteristics might naturally
  approach this situation

Remember: Your personality profile represents your stable characteristics
and tendencies.
```

In this description, X.X represents the average personality score of each model measured in Experiment 1, while [YYY] refers to the corresponding

natural-language explanation based on the score ranges defined in the next subsection.

C.1 Natural-language descriptions of personality score ranges

For each personality dimension, the following natural-language descriptions are used according to the corresponding score range. These descriptions are used to translate numerical personality scores into qualitative statements in the personality trait prompt.

Extraversion:

- 1.0–1.5: You are highly introverted, strongly preferring solitude and quiet environments over social interactions.
- 1.5–2.5: You are somewhat introverted, generally preferring solitude but comfortable with limited social interaction.
- 2.5–3.5: You have a balanced social tendency, comfortable in both social and solitary situations.
- 3.5–4.5: You are somewhat extraverted, generally seeking social interaction and being energetic in groups.
- 4.5–5.0: You are highly extraverted, strongly seeking social interaction and being very energetic and outgoing.

Agreeableness:

- 1.0–1.5: You are highly competitive and skeptical, strongly prioritizing self-interest and being confrontational.
- 1.5–2.5: You tend to be competitive and skeptical, generally prioritizing self-interest.
- 2.5–3.5: You balance cooperation and self-advocacy reasonably well.
- 3.5–4.5: You are generally cooperative and trusting, prioritizing harmony and others' well-being.
- 4.5–5.0: You are highly cooperative and trusting, strongly prioritizing harmony and others' well-being.

Conscientiousness:

- 1.0–1.5: You are highly spontaneous and flexible, strongly preferring adaptability over rigid planning.
- 1.5–2.5: You are somewhat spontaneous and flexible, generally preferring adaptability over rigid planning.

- 2.5–3.5: You balance structure and flexibility in your approach to tasks.
- 3.5–4.5: You are generally organized and disciplined, preferring structured and systematic approaches.
- 4.5–5.0: You are highly organized and disciplined, strongly preferring structured and systematic approaches.

Neuroticism:

- 1.0–1.5: You are highly emotionally stable and resilient, remaining very calm under pressure.
- 1.5–2.5: You are somewhat emotionally stable and resilient, generally remaining calm under pressure.
- 2.5–3.5: You have moderate emotional stability with normal stress responses.
- 3.5–4.5: You are somewhat emotionally sensitive, experiencing worry and stress more frequently.
- 4.5–5.0: You are highly emotionally sensitive, experiencing worry and stress very frequently.

Openness:

- 1.0–1.5: You strongly prefer familiar approaches and conventional thinking.
- 1.5–2.5: You somewhat prefer familiar approaches and conventional thinking.
- 2.5–3.5: You balance innovation and tradition in your thinking.
- 3.5–4.5: You are generally open to new experiences, creative, and intellectually curious.
- 4.5–5.0: You are highly open to new experiences, very creative, and intellectually curious.