Human-level play in the game of Diplomacy by combining language models with strategic reasoning

Yangjae Lee
RL Paper Study
2023/04/03

FAIR Diplomacy Team


Anton Bakhtin


Noam Brown


Emily Dinan


Colin Flaherty


Jonathan Gray


Hengyuan Hu


Athul Paul Jacob


Adam Lerer


Mike Lewis


Alexander Miller

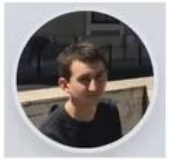

Adithya Renduchintala


Weiyan Shi


David Wu


Hugh Zhang


Gabriele Farina


Daniel Fried


Andrew Goff


Mojtaba Komeili


Minae Kwon


Karthik Konath


Sasha Mitts


Stephen Roller


Dirk Rowe

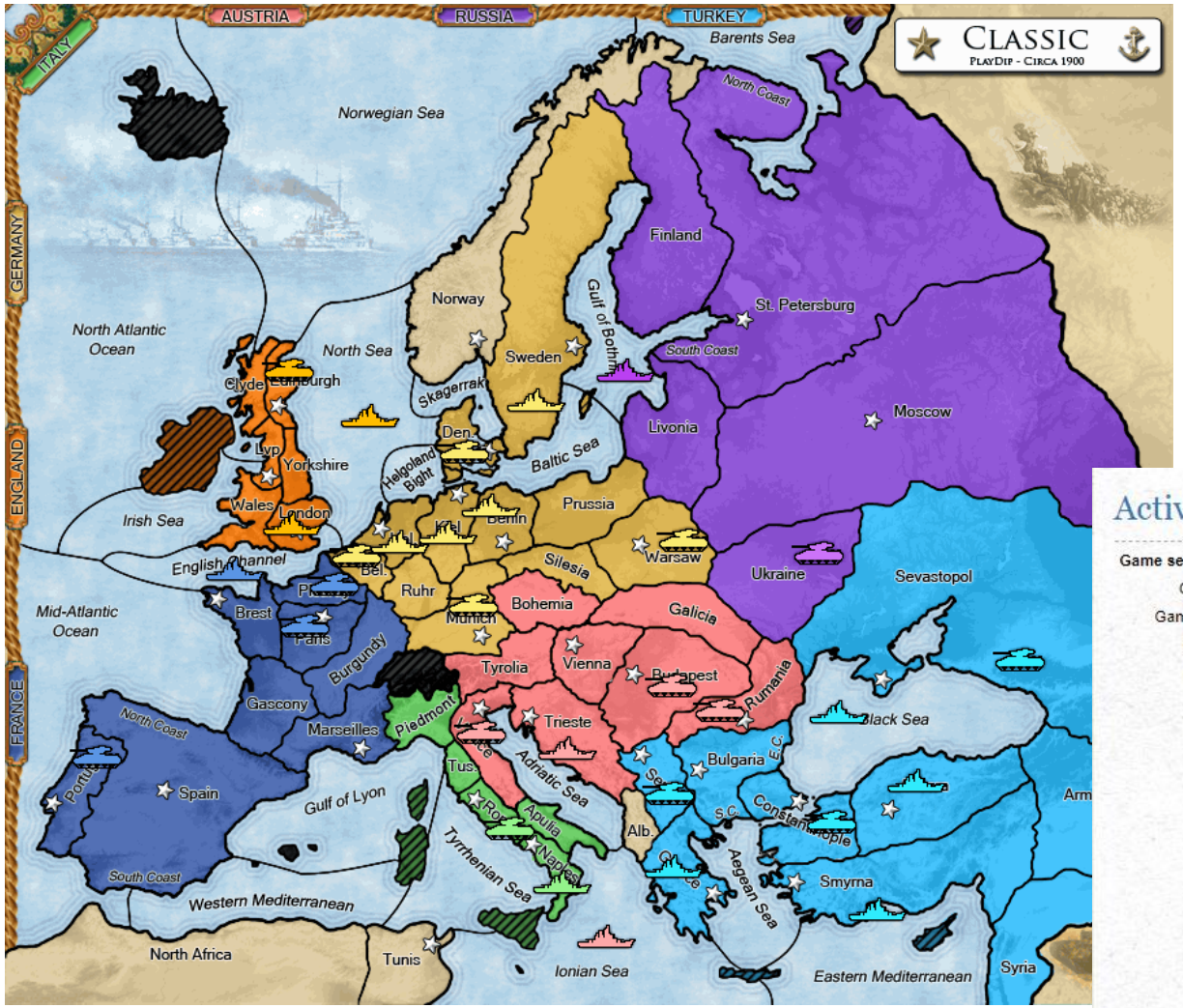

Joe Spisak


Alex Wei

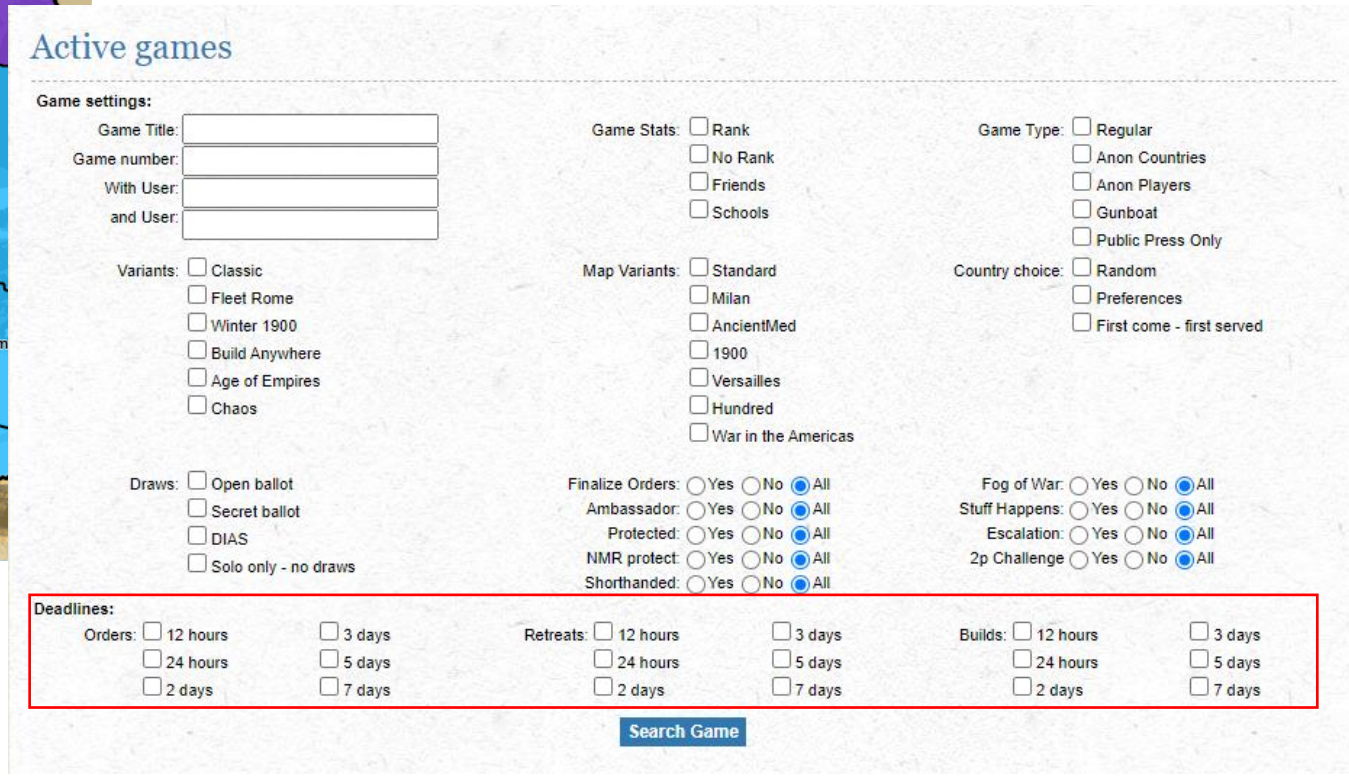

Markus Zijlstra

NLP, Game Theory, RL 외 다양한 분야 전문가들의 협업

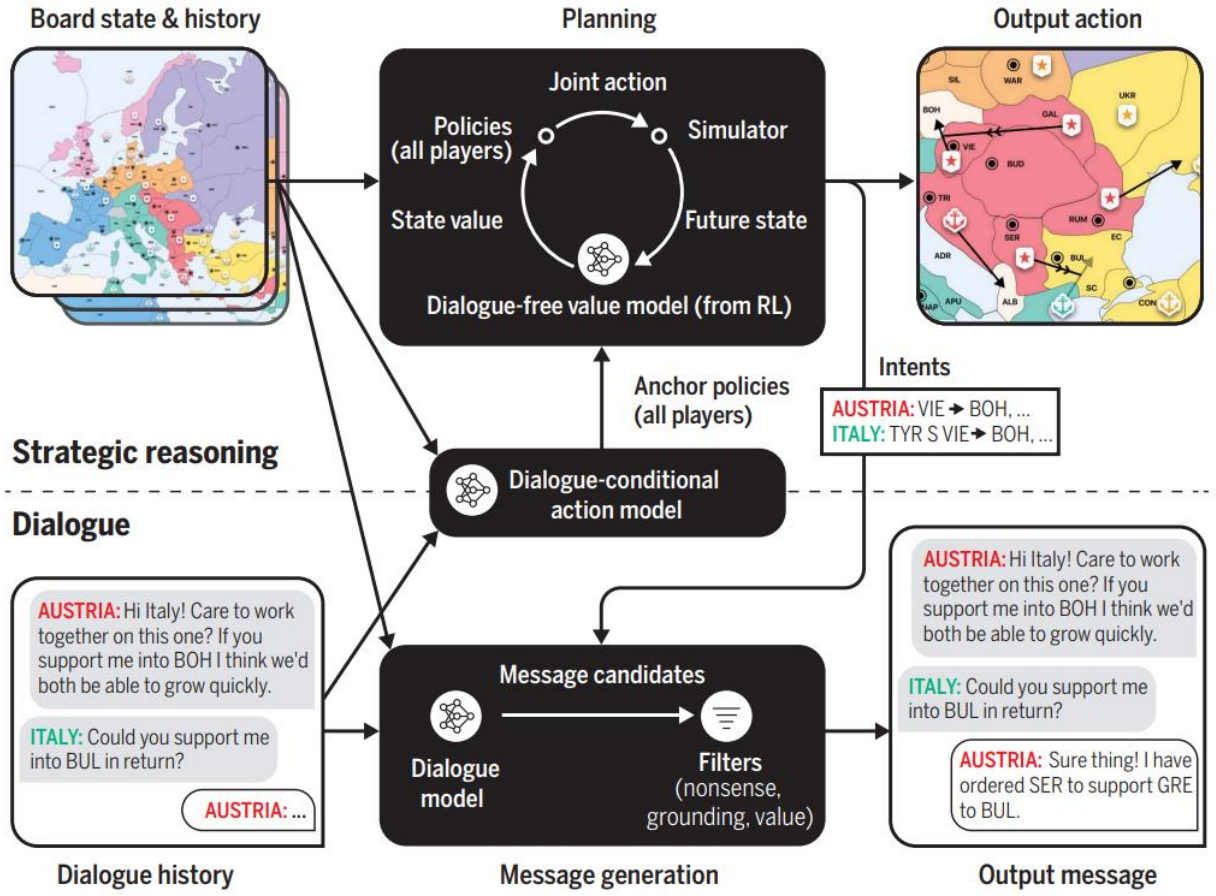
Game 설명



- 게임은 다음 맵에서 7명이 각자의 지역으로 시작
- 1:1 private한 대화가 가능하여 협업 가능
- 보유한 유닛들을 움직여 땅을 따먹고, 모든 땅을 점령하거나 마지막 한명이 남으면 승리
- 한 턴에 최소 12시간에서 7일까지 소요



Overview



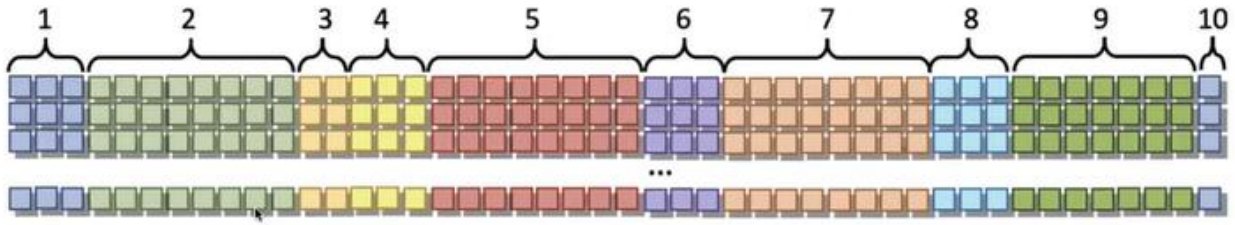
- strategic planning engine, dialogue agent 두개의 영역으로 구분
- Planning engine: cicero가 좋은 성과를 낼 수 있도록 상대방과 cicero의 행동을 어떻게 할지 추론
- dialogue agent: planning engine에서 cicero가 도출한 결과를 다른 게임 플레이어에게 전달하기 위해 대화로 변환하거나 다른 플레이어의 대화를 변환하여 의도 파악

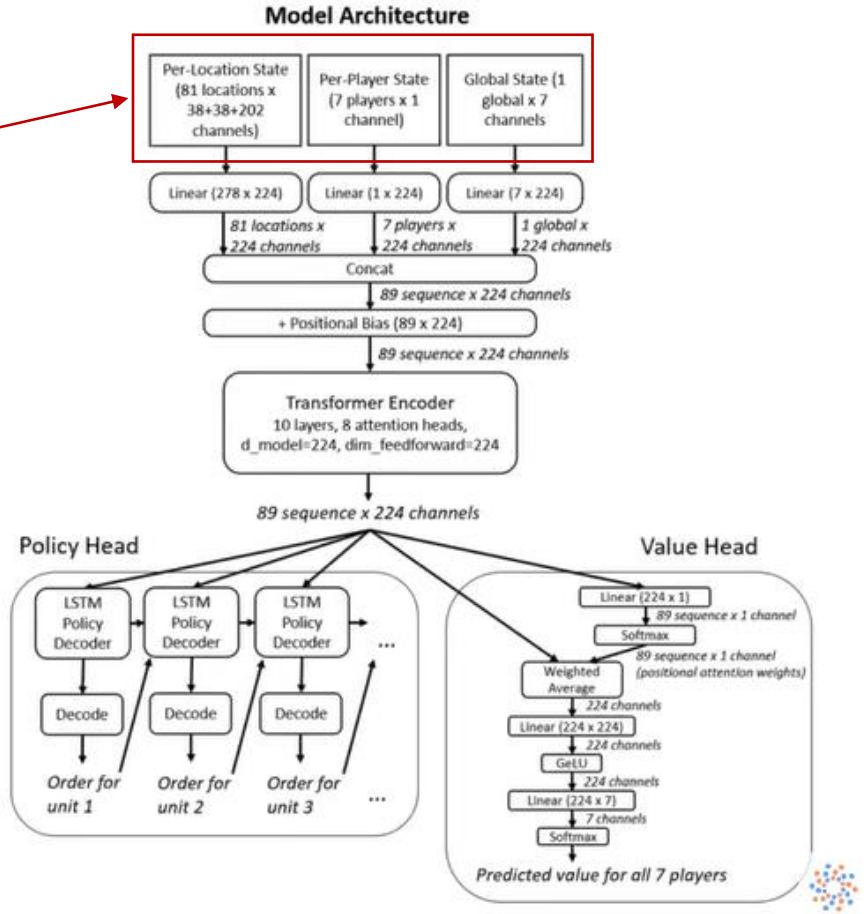
MDP, Network

Modeling Human Policies

Board State Encoding

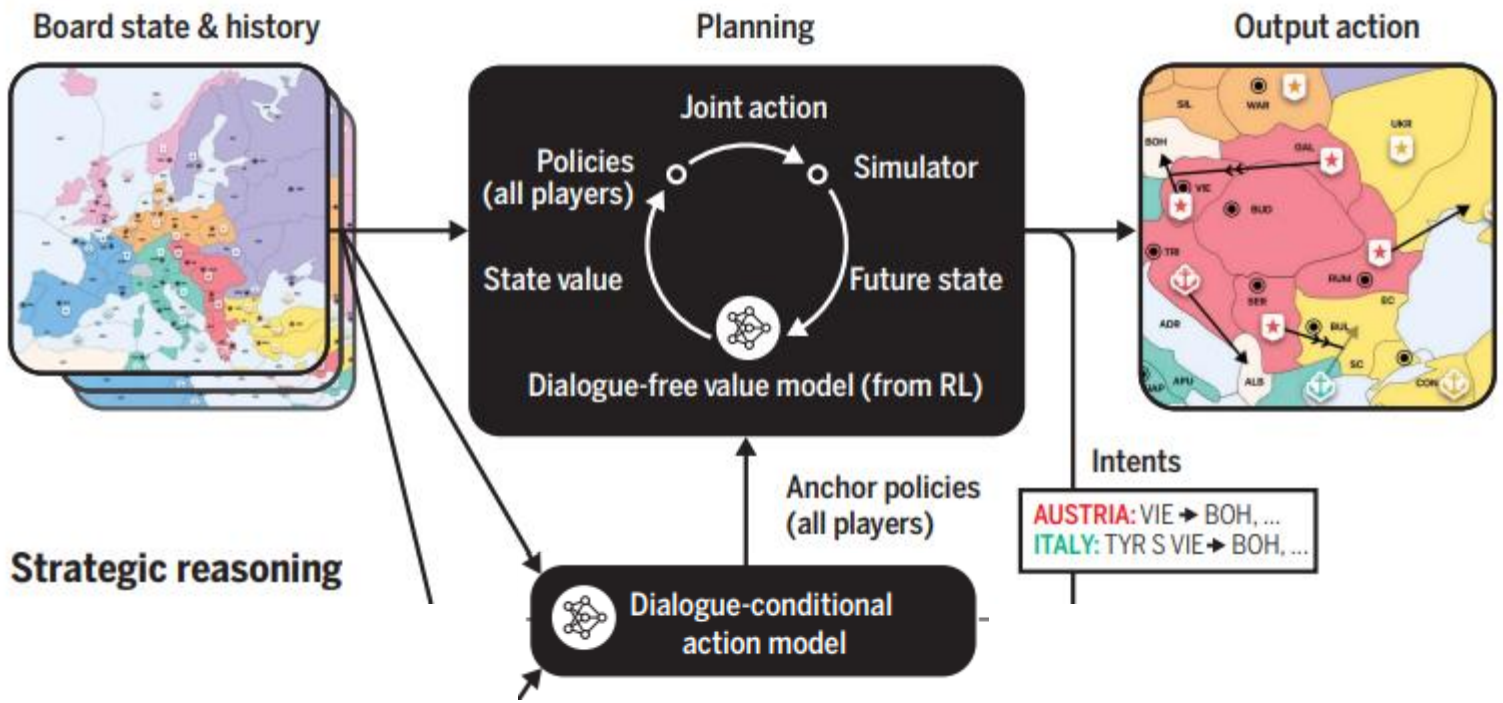
1. Unit Type (Army, Fleet, None)
2. Unit Power (AUS, ENG, FRA, ITA, GER, RUS, TUR, None)
3. Buildable, Removable
4. Dislodged Unit Type (Army Fleet, None)
5. Dislodged Unit Power (AUS, ..., TUR, None)
6. Area Type (Land, Water, Coast)
7. Supply Center Owner (AUS, ..., TUR, None)
8. *Season (Spring, Fall, Winter)*
9. *Build Numbers (Integer value for each power)*
10. *Press Type (NoPress, Press)*





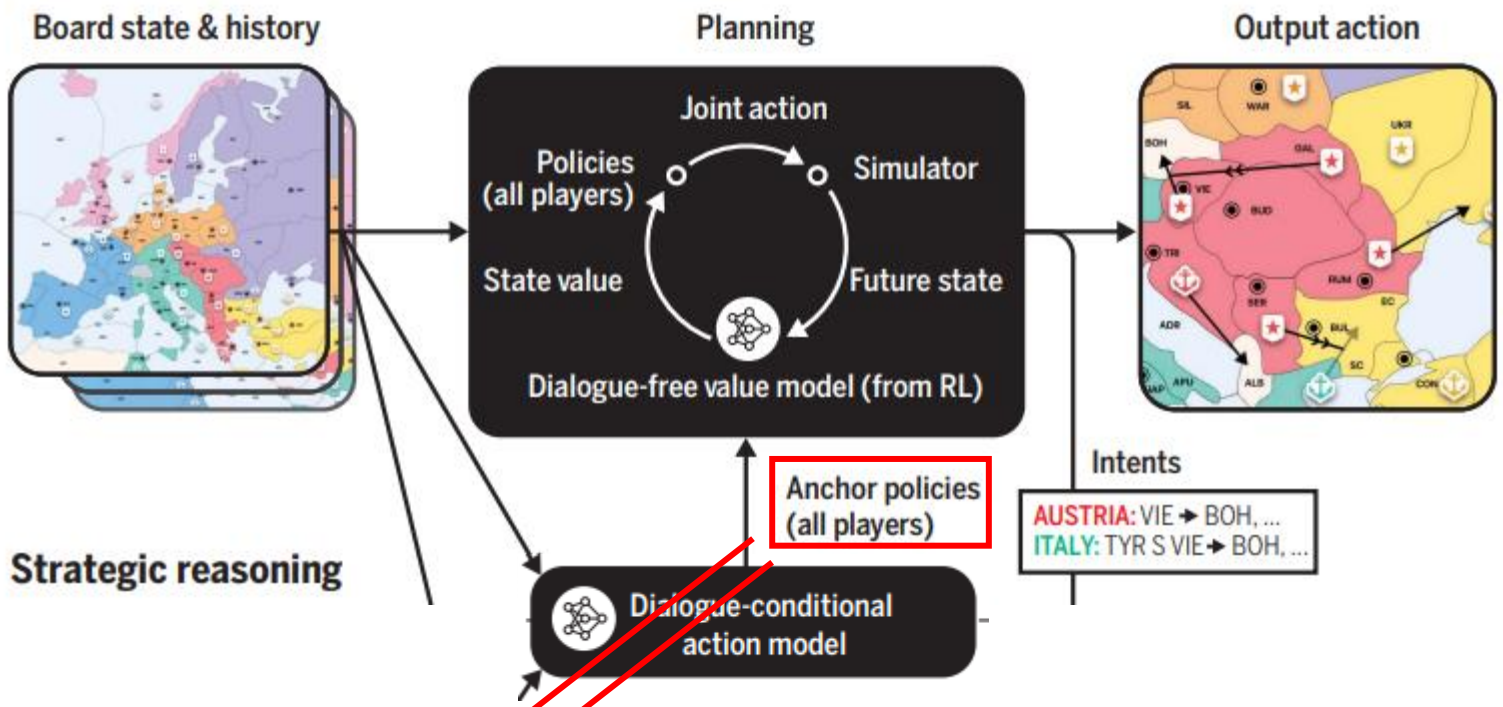
action: 유닛 이동

Strategic Planning Engine - 1



1. 다른 플레이어의 policy 추측
   - AlphaGo의 MTCS가 1:1이 아니라 n으로 늘어났다고 보시면 됩니다.
   - 해당 플레이어가 expected value를 높이기 위해 어떤 행동을 할 것인가를 추측합니다.
2. joint action은 cicero가 하는 action과는 조금 다름. 모든 플레이어의 행동(a1, a2, ...)들이 합쳐져 joint action a 형성
3. 시뮬레이터에 joint action을 넣고 loop를 반복함으로써 planning engine의 성능을 향상시킴

Strategic Planning Engine - 2

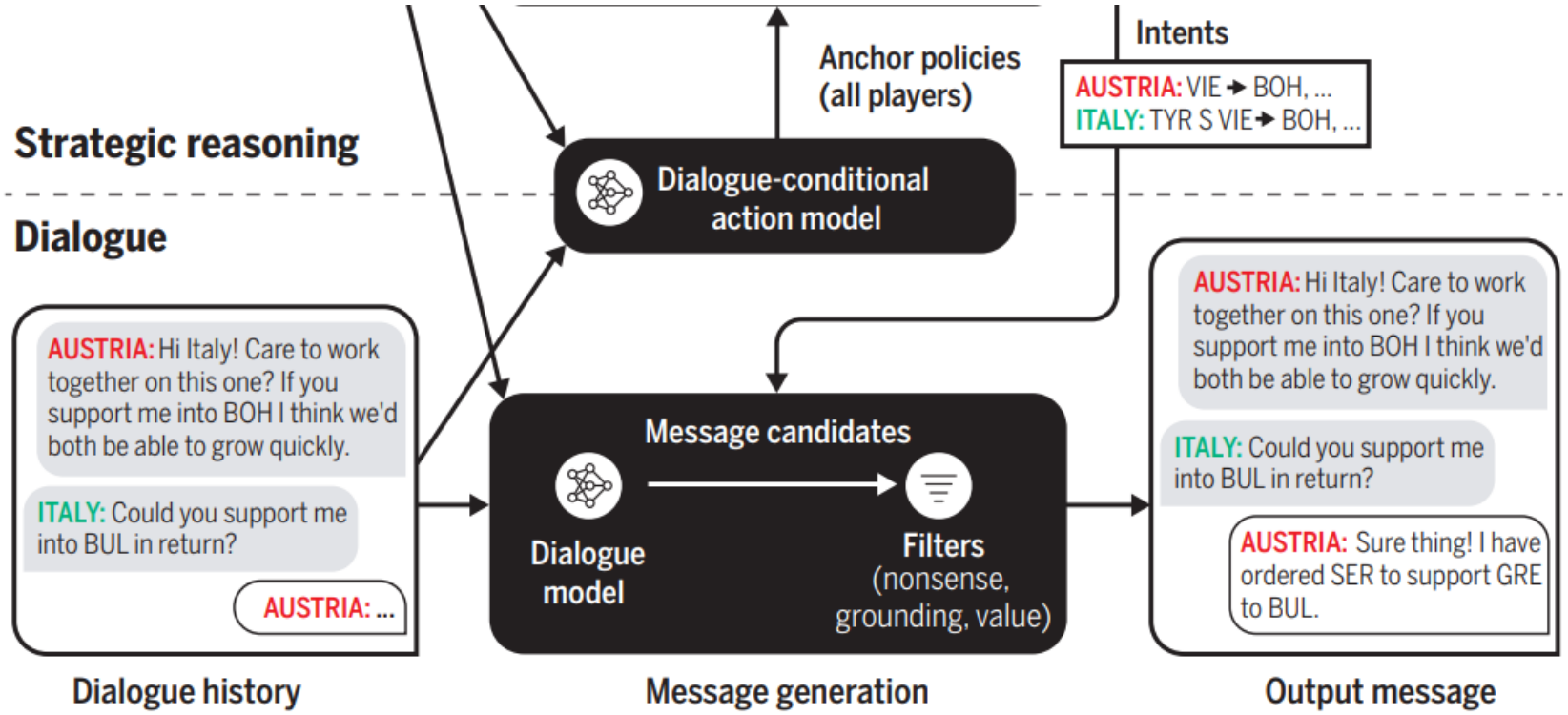


논문에는 해당 식은 없음. 저자가 발표한 자료에서 발췌

$$L(\pi) = -V(\pi) + \lambda\, KL(\pi || \pi_{human})$$

1. Anchor policy
   - Imitation Learning policy
   - 강화학습만으로 학습하

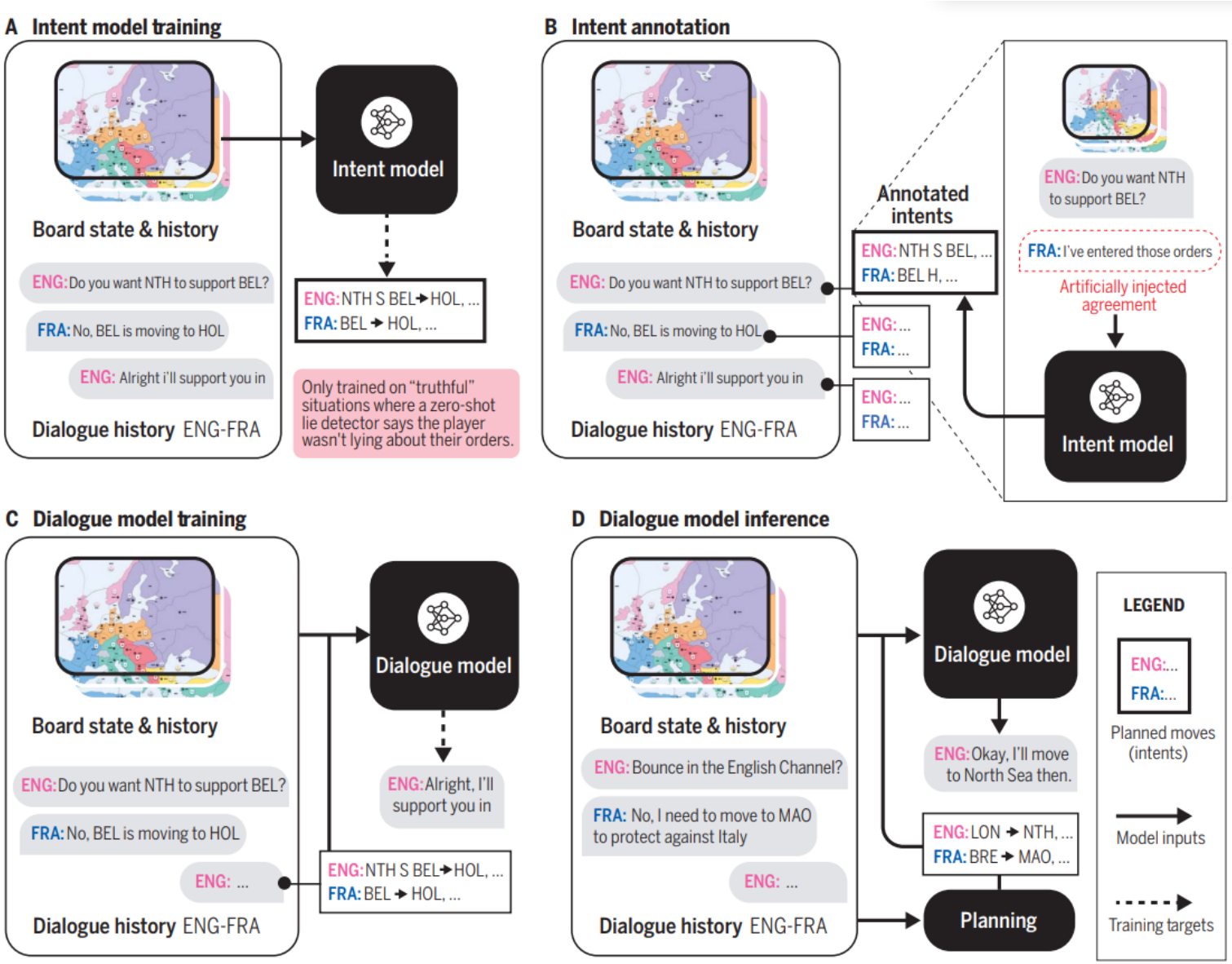
Dialogue Agent - 1



1. Intent
   - Planning engine에서 생성된 모든 플레이어들의 목표와 의도를 나타냄

2. LM(BART)
   - Game state, 대화내용, intents를 바탕으로 문장 생성

Dialogue Agent - 2



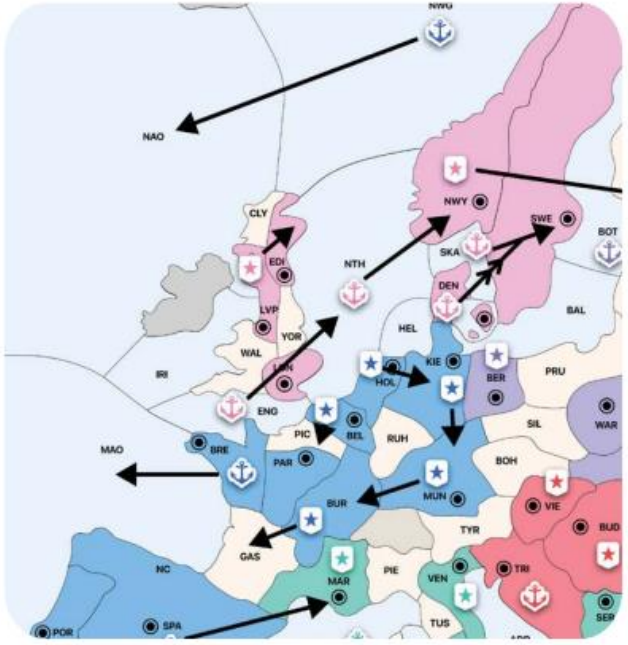
1. Intent model
   - 초반엔 전문가가 annotation 으로 학습
   - 이후에 학습된 모델을 바탕으로 더 큰 데이터셋을 학습

2. 대화는 planning모델에서 의도 파악 후 대화와 같이 dialogue model에 들어가 문장 생성

Dialogue Agent - 3

England agrees:

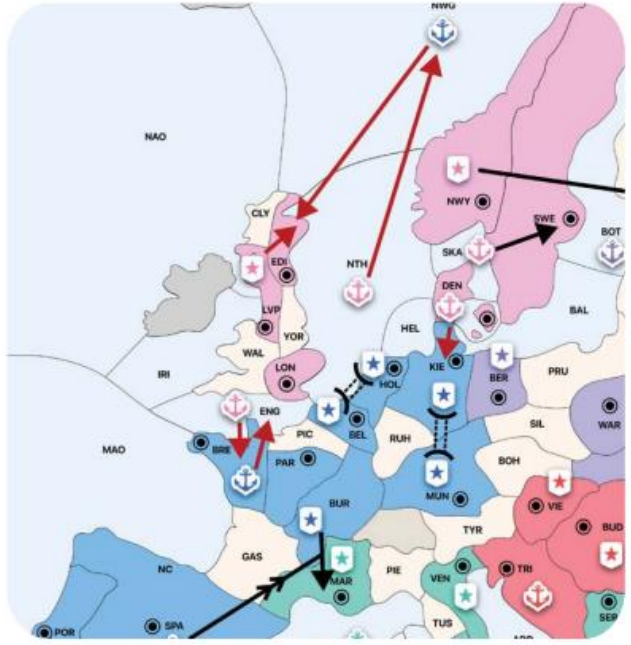
ENG → FRA Yes! I will move out of ENG if you head back to NAO.



Cicero predicts England will retreat from ENG to NTH 85% of the time, backs off its own fleet to NAO as agreed, and begins to move armies away from the coast.

England is hostile:

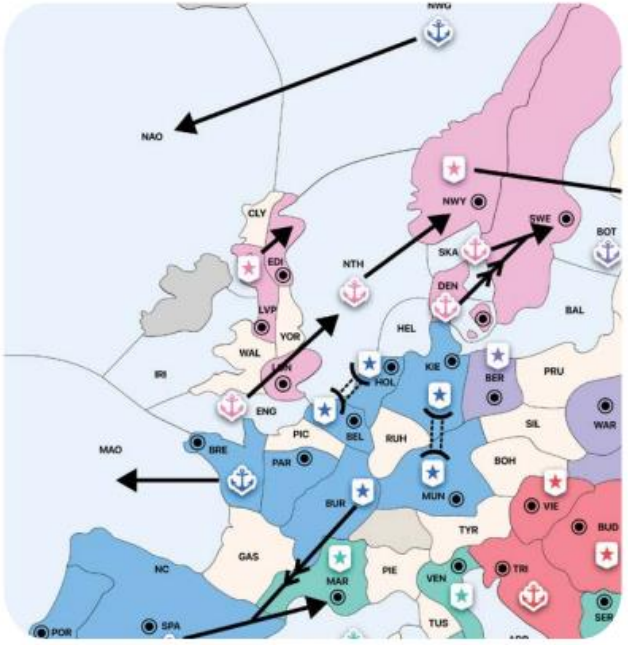
ENG → FRA You've been fighting me all game. Sorry, I can't trust that you won't stab me.



Cicero does not back off its fleet but rather attacks EDI with it, and leaves its armies at the coast to defend against an attack from England, predicting that England will attack about 90% of the time.

England tries to take advantage of Cicero:

ENG → FRA Yes! I'll leave ENG if you move KIE -> MUN and HOL -> BEL.



Strategic planning rejects the possibility of vacating KIE and HOL, because it would make Cicero too vulnerable. Cicero backs off its fleet to NAO but keeps armies at the coast to defend.

1. State가 모두 동일할 때, 상대방의 대화에서 의도를 파악하여 그에 따른 전략 수립

Filtering

1. Dialogue model이 이전 대화와 모순되거나 현재 게임상태와 일치하지 않은 메시지 생성
2. 불가능한 계획에 대한 논의 또는 비정상적인 움직임에 대한 메시지 생성
3. 전략 유출과 같은 취약한 정보를 적에게 유출

Ensemble of Classifiers

- Intent correspondence
- Value-based filtering
- Counterfactual classifier

1. Classifier를 통해 오류가 포함된 메시지 필터링
2. intent의 가능성이 어떻게 변화할지 알려주는 Metric 계산을 통해 이상 행동 탐지
   - 메시지를 보내기 전, 메시지를 보낸 것처럼 가정하고 행동 후 일정 수치 이하면 필터링
3. 메시지에 점수를 부여하여 주요 정보 유출 방지
   - 메시지를 보냈을 때 상대 플레이어의 policy에 끼치는 영향력을 측정하여 필터링

Results

1. 게임 플레이 성능
   - 1판이상 플레이한 플레이어 기준 상위 10%의 실력
   - 하지만 여전히 전문가가 보기에는 부족한 부분이 많다.
2. 대화 성능
   - BART를 게임에서 많이 쓰는 문장과 언어로 fine tuning
   - 같이 플레이한 플레이어가 알아차리지 못할 정도의 자연스러움을 보여줌