

SKRIPSI

**PEMODELAN LAWAN BELAJAR MENGGUNAKAN NONLINEAR
AUTOREGRESSIVE WITH EXOGENOUS INPUT (NARX) PADA
ITERATED PRISONER'S DILEMMA DENGAN HORIZON TERBATAS**

**LEARNING OPPONENT MODELLING USING A NONLINEAR
AUTOREGRESSIVE WITH EXOGENOUS INPUT (NARX) MODEL IN THE
ITERATED PRISONER'S DILEMMA WITH A BOUNDED HORIZON**



FAQIH MAHARDIKA
21/482551/PA/21039

**PROGRAM STUDI ILMU KOMPUTER
DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS GADJAH MADA
YOGYAKARTA**

2026

SKRIPSI

PEMODELAN LAWAN BELAJAR MENGGUNAKAN NONLINEAR AUTOREGRESSIVE WITH EXOGENOUS INPUT (NARX) PADA ITERATED PRISONER'S DILEMMA DENGAN HORIZON TERBATAS

LEARNING OPPONENT MODELLING USING A NONLINEAR AUTOREGRESSIVE WITH EXOGENOUS INPUT (NARX) MODEL IN THE ITERATED PRISONER'S DILEMMA WITH A BOUNDED HORIZON

Diajukan untuk memenuhi salah satu syarat memperoleh derajat
Sarjana Sains Ilmu Komputer



FAQIH MAHARDIKA
21/482551/PA/21039

PROGRAM STUDI ILMU KOMPUTER
DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS GADJAH MADA
YOGYAKARTA

2026

HALAMAN PENGESAHAN

SKRIPSI

PEMODELAN LAWAN BELAJAR MENGGUNAKAN NONLINEAR AUTOREGRESSIVE WITH EXOGENOUS INPUT (NARX) PADA ITERATED PRISONER'S DILEMMA DENGAN HORIZON TERBATAS

Telah dipersiapkan dan disusun oleh

FAQIH MAHARDIKA
21/482551/PA/21039

Telah dipertahankan di depan Tim Penguji
pada tanggal 8 Mei 2026

Susunan Tim Penguji

Dr. Sri Mulyana, M.Kom.
Pembimbing

Bob, M.Sc.
Ketua Penguji

Eve, Ph.D.
Anggota Penguji

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 8 Mei 2026

Faqih Mahardika

DAFTAR ISI

Halaman Judul	ii
Halaman Pengesahan	iii
Halaman Pernyataan	iv
DAFTAR ISI	v
DAFTAR TABEL	viii
DAFTAR GAMBAR	ix
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Penelitian	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
1.6 Sistematika Penelitian	5
II TINJAUAN PUSTAKA	6
2.1 Synthesis / Discussion	6
2.1.1 Asumsi perilaku lawan dalam <i>opponent modelling</i> untuk <i>Repeated Games</i>	6
2.1.2 Pendekatan metodologis dalam <i>opponent modelling</i> untuk <i>Repeated Games</i>	7
2.1.3 Bagaimana efektivitas strategi <i>opponent modelling</i> dievaluasi dalam <i>Repeated Games</i> ?	9
2.2 Kesimpulan	11
III LANDASAN TEORI	16
3.1 Dilema Sosial	16
3.2 Game Theory	16
3.3 Prisoner's Dilemma	17
3.4 Iterated Prisoner's Dilemma	17

3.4.1	Finite Horizon	17
3.4.2	Infinite Horizon dan Discount Factor	18
3.4.3	Stochastic Termination	18
3.4.4	Hubungan Stochastic Termination dan Discounted Return	19
3.4.5	Definisi History Interaksi pada mekanisme online	19
3.4.6	Incomplete Information dan Pembentukan Belief	20
3.5	Pemodelan Lawan	20
3.6	RNN	21
3.6.1	Recurrent Neural Network untuk Opponent Modelling	21
3.6.2	Latent Belief Representation	22
3.7	Long Short-Term Memory (LSTM)	22
3.7.1	Input Gate	23
3.7.2	Candidate Cell State	23
3.7.3	Forget Gate	23
3.7.4	Cell State (Memori Jangka Panjang)	24
3.7.5	Output Gate	24
3.7.6	Hidden State Update	24
3.8	Prediksi Multi-Langkah	25
3.8.1	Forecasting Rekursif dan Monte Carlo Rollout	25
3.8.2	Dari Akurasi Prediksi ke Kinerja Strategis	25
3.8.3	Decoupled Opponent Modelling	28
3.8.4	Risiko Propagasi Kesalahan pada Forecasting Rekursif	29
3.8.5	Evaluasi Multi-Step Menggunakan Negative Log-Likelihood .	29
IV ANALISIS DAN PERANCANGAN		31
4.1	Deskripsi Umum	31
4.2	Formulasi Masalah	31
4.3	Metodologi Penelitian	31
4.3.1	Formalisasi Permainan	32
4.3.2	Arsitektur Model	32
4.3.3	Fungsi Objektif	34
4.3.4	Simulasi Monte Carlo Rollout	34
4.3.5	Pengumpulan Data	35
4.3.6	Evaluasi	36
V HASIL DAN PEMBAHASAN		37

VI KESIMPULAN DAN SARAN	38
DAFTAR PUSTAKA	39
LAMPIRAN	42

DAFTAR TABEL

2.1	Ringkasan Penelitian Terkait	13
2.2	Asumsi perilaku lawan berdasarkan dependensi perilaku yang dapat diamati.	14
2.3	Lingkungan evaluasi dan metrik dalam penelitian	15

DAFTAR GAMBAR

INTISARI

PEMODELAN LAWAN BELAJAR MENGGUNAKAN NONLINEAR AUTOREGRESSIVE WITH EXOGENOUS INPUT (NARX) PADA ITERATED PRISONER'S DILEMMA DENGAN HORIZON TERBATAS

oleh

Faqih Mahardika

21/482551/PA/21039

Kerja sama dan konflik merupakan karakteristik fundamental dari interaksi dalam masyarakat, sistem biologis, dan lingkungan agen-artifisial, di mana agen secara berulang menghadapi pertukaran strategis antara kepentingan diri jangka pendek ataupun hasil kolektif jangka panjang. *Opponent Modelling* memiliki peran dasar dalam konteks tersebut dengan memungkinkan agen untuk menginferensi, mengantisipasi, dan beradaptasi terhadap perilaku pihak lain, dengan pengaplikasiannya mencakup negosiasi, interaksi pasar, hingga sistem kecerdasan buatan multi-agenn. Meskipun informatif untuk kinerja jangka panjang, metrik-metrik tersebut umumnya diterapkan pada horizon interaksi yang tidak dibatasi atau cukup panjang, sehingga membatasi pemahaman mengenai efisiensi dan ketepatan waktu dalam proses identifikasi serta adaptasi terhadap lawan secara langsung. Tinjauan ini menemukan adanya kesenjangan struktural dalam praktik evaluasi yang ada dan menekankan perlunya kerangka penilaian yang sadar akan horizon (*horizon-aware*) agar lebih merefleksikan keterbatasan interaksi berulang di dunia nyata.

Cooperation and conflict are fundamental features of interaction in human societies, biological systems, and artificial-agent environments, where agents repeatedly face strategic trade-offs between short-term self-interest and long-term collective outcomes. Opponent modelling plays a central role in such settings by enabling agents to infer, anticipate, and adapt to the behavior of others, with applications ranging from negotiation and market interactions to multi-agent artificial intelligence systems. This literature review synthesizes prior work on opponent modelling in repeated strategic games, using the Iterated Prisoner’s Dilemma as a canonical instantiation of repeated social dilemmas rather than as a restrictive domain. The review analyzes existing approaches along three dimensions: assumptions about opponent behavior, opponent-modelling methodologies, and evaluation practices. The surveyed literature exhibits substantial diversity in opponent behavior assumptions, including stationary, reactive, learning-based, and population-mediated opponents, often embedded implicitly within experimental setups. Methodologically, approaches span gradient-based learning, deep reinforcement learning, recursive belief reasoning, evolutionary dynamics, and system identification. Evaluation practices predominantly emphasize outcome-based metrics such as cooperation rate, average payoff, and equilibrium-related measures. While informative for long-horizon performance, these metrics are typically applied under unconstrained interaction horizons, limiting insight into the efficiency and timeliness of opponent identification and adaptation. This review highlights a structural gap in existing evaluations and underscores the need for horizon-aware assessment frameworks that better reflect the constraints of real-world repeated interactions.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Interaksi strategis berulang antara agen rasional merupakan fondasi penting dalam multi-agent reinforcement learning (MARL) (Hernandez-Leal et al. 2019). Salah satu kerangka klasik yang sering digunakan untuk mempelajari dinamika adaptasi strategis adalah *Repeated Iterated Prisoner's Dilemma* (IPD). Dalam permainan ini, dua agen berinteraksi secara berulang tanpa komunikasi eksplisit, dan negosiasi strategi terjadi murni melalui urutan aksi(Axelrod dan Hamilton 1981).

Pada lingkungan IPD dengan *stochastic termination* (geometric stopping), dinamika menjadi lebih kompleks karena horizon interaksi bersifat tidak pasti (Axelrod dan Hamilton 1981). Kompleksitas ini meningkat ketika lawan bersifat adaptif dan non-stasioner, misalnya menggunakan algoritma pembelajaran daring seperti *Hedge*, *Online Mirror Descent*, atau pendekatan berbasis reinforcement learning. Dalam konteks ini, agen tidak mengetahui algoritma internal lawan dan harus beradaptasi hanya berdasarkan observasi historis (Nisan et al. 2008).

Dalam konteks terminasi stokastik dengan distribusi geometrik, horizon efektif interaksi bersifat terbatas secara ekspektasi. Oleh karena itu, pemodelan prediksi jangka pendek menjadi lebih relevan dibandingkan analisis asimtotik jangka panjang, karena keputusan eksplorasi harus mempertimbangkan probabilitas terminasi yang meningkat seiring waktu (Nisan et al. 2008).

Dalam lingkungan dengan horizon stokastik dan lawan non-stasioner, sinyal reward seringkali menjadi indikator yang terkonflisi antara kualitas kebijakan agen dan perubahan strategi lawan(**kasih referensi**). Ketika reward digunakan secara langsung untuk memperbarui model lawan, terjadi risiko pencampuran antara inferensi dinamika lawan dan evaluasi utilitas tindakan. Akibatnya, proses pembelajaran dapat mengalami bias identifikasi dan ketidakstabilan adaptasi (**kasih referensi**). Hingga saat ini, belum banyak pendekatan yang secara eksplisit memisahkan pemodelan belief terhadap lawan dari optimisasi reward, sekaligus mengevaluasi perolehan informasi epistemik jangka pendek dalam interaksi strategis berulang dengan terminasi stokastik.

Sebagian besar pendekatan opponent modelling dalam MARL berfokus pa-

da inferensi tipe eksplisit, estimasi utilitas, atau pelatihan bersama berbasis reward. Pendekatan tersebut umumnya:

1. Mengintegrasikan belief modelling langsung ke dalam objective reward,
2. Berorientasi pada prediksi satu langkah ke depan,
3. Tidak memisahkan modul belief dari optimisasi kebijakan,
4. Tidak mengevaluasi *epistemic information gain* secara eksplisit.

Akibatnya, eksplorasi terhadap lawan sering kali bersifat implisit dan tidak terstruktur sehingga kurang robust terhadap perubahan algoritma maupun dinamika pembelajaran lawan.**(kasih referensi)**

Dalam konteks ini, ketidakpastian terhadap dinamika strategi lawan bukan sekadar noise statistik, melainkan komponen epistemik yang secara langsung memengaruhi kualitas pengambilan keputusan. Oleh karena itu, eksplorasi tidak hanya berkaitan dengan perolehan reward yang lebih tinggi, tetapi juga dengan reduksi ketidakpastian terhadap model lawan. Pendekatan yang tidak secara eksplisit mengkantifikasi aspek epistemik ini berpotensi gagal beradaptasi ketika lawan mengubah kebijakannya secara dinamis.

Berdasarkan permasalahan tersebut, diperlukan suatu kerangka pemodelan yang (i) tidak mengasumsikan struktur parametrik eksplisit dari algoritma lawan, (ii) memisahkan secara konseptual dan komputasional belief modelling dari optimisasi reward, serta (iii) memungkinkan integrasi estimasi perolehan informasi epistemik ke dalam mekanisme seleksi aksi.

Dalam kerangka yang diusulkan, sinyal reward tetap digunakan untuk evaluasi dan seleksi kebijakan, namun tidak digunakan sebagai sinyal pelatihan dalam pembaruan belief terhadap lawan. Dengan demikian, proses inferensi dinamika lawan dipisahkan secara konseptual dari optimisasi utilitas agen.

Dengan pendekatan ini, belief terhadap dinamika lawan tidak hanya berfungsi sebagai model prediktif, tetapi menjadi komponen aktif dalam menentukan strategi eksplorasi dan eksloitasi agen.

Dalam penelitian ini, peningkatan keyakinan (confidence improvement) didefinisikan sebagai perubahan terukur pada distribusi prediktif belief akibat observasi kontrafaktual trajectory aksi agen, yang berfungsi sebagai proksi reduksi ketidakpastian epistemik jangka pendek.

Sebagai ilustrasi, estimasi tersebut dapat diintegrasikan ke dalam skema seleksi aksi berbasis prinsip seperti Upper Confidence Bound (UCB), meskipun perancangan mekanisme seleksi aksi bukan fokus utama penelitian ini.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, penelitian ini merumuskan permasalahan sebagai berikut:

1. Bagaimana membangun modul pemodelan lawan yang sepenuhnya observasional dan tidak bergantung pada sinyal reward?
2. Bagaimana melakukan prediksi multi-langkah jangka pendek (k -step forecasting) terhadap distribusi aksi lawan yang dikondisikan pada trajectory aksi agen?
3. Bagaimana mengestimasi peningkatan keyakinan (*confidence improvement*) yang didefinisikan sebagai perubahan distribusi prediktif belief

1.3 Batasan Penelitian

Agar fokus penelitian tetap terjaga dan implementasi eksperimental stabil, penelitian ini dibatasi pada:

1. Lingkungan *Repeated Iterated Prisoner's Dilemma* dengan *stochastic termination*.
2. Lawan adaptif yang tidak diketahui algoritmanya, namun terbatas pada kelas pembelajaran daring seperti Hedge, Online Mirror Descent, serta pendekatan reinforcement learning berbasis policy update tanpa observasi struktur eksplisit lawan.
3. Prediksi jangka pendek (k -step forecasting), tanpa pemodelan asimtotik jangka panjang.
4. Tidak mencakup inferensi eksplisit terhadap struktur algoritma lawan.
5. Tidak mengoptimalkan mutual information atau objective identifiability struktural secara eksplisit, melainkan menggunakan proksi berbasis perubahan distribusi prediktif belief.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mengembangkan modul *decoupled latent opponent belief* yang dilatih murni menggunakan forecasting loss tanpa sinyal reward.
2. Mengimplementasikan mekanisme prediksi multi-langkah jangka pendek terhadap distribusi aksi lawan yang dikondisikan pada trajectory aksi agen.
3. Merumuskan ukuran *counterfactual confidence improvement* sebagai proksi reduksi ketidakpastian prediktif jangka pendek.
4. Mengeksplorasi bagaimana komponen belief dan estimasi reduksi ketidakpastian dapat dimanfaatkan dalam mekanisme seleksi aksi yang mempertimbangkan eksplorasi dan eksplorasi.

1.5 Manfaat Penelitian

Kontribusi utama penelitian ini adalah sebagai berikut:

1. Mengusulkan modul belief lawan yang terpisah dari optimisasi reward tanpa asumsi struktur parametrik eksplisit terhadap algoritma lawan.
2. Memperkenalkan pendekatan *multi-step short-horizon forecasting* dalam konteks IPD adaptif.
3. Mengusulkan estimasi *counterfactual confidence improvement* sebagai ukuran eksplorasi epistemik berbasis prediksi.
4. Memberikan formulasi eksplorasi epistemik berbasis pembaruan belief kontrafaktual yang dapat diaplikasikan pada lingkungan interaksi strategis non-stasioner.
5. Memberikan kerangka eksplorasi epistemik jangka pendek pada interaksi strategis berulang non-stasioner yang tidak bergantung pada asumsi struktur algoritma lawan.

1.6 Sistematika Penelitian

1. Bab 1: Pendahuluan: Latar belakang, rumusan masalah, batasan, tujuan, manfaat, dan sistematika penelitian.
2. Bab 2: Tinjauan Pustaka: Kajian literatur terkait opponent modelling, IPD, dan pendekatan pembelajaran daring.
3. Bab 3: Kerangka Teoritis: Formulasi masalah, definisi metrik, dan arsitektur umum.
4. Bab 4: Metode Penelitian: Desain eksperimen, algoritma, dan prosedur evaluasi.
5. Bab 5: Jadwal Penelitian: Rencana waktu pelaksanaan penelitian.

BAB II

TINJAUAN PUSTAKA

sebanyak 13 studi dimasukkan ke dalam tinjauan sistematis final. Judul dan karakteristik dari studi-studi tersebut disajikan pada Tabel ??.

2.1 Synthesis / Discussion

Bagian ini mensintesis temuan-temuan utama dari studi yang ditinjau dengan tujuan mengidentifikasi pola asumsi perilaku lawan, keterbatasan metodologis yang berulang, serta celah penelitian yang masih terbuka dalam *opponent modelling* untuk *repeated games*. Sintesis difokuskan pada dimensi perilaku yang dapat diamati dari riwayat interaksi, alih-alih pada asumsi internal seperti tujuan optimisasi, struktur pembaruan parameter, atau representasi kebijakan lawan yang sering kali tidak dapat diakses secara langsung oleh agen.

Meskipun berbagai atribut telah diidentifikasi dalam studi-studi yang disertakan, hanya dimensi yang bersifat diskriminatif secara metodologis yang digunakan untuk perbandingan lintas karya. Atribut terkait lingkungan permainan dan protokol evaluasi dibahas secara terpisah untuk menghindari pencampuran antara asumsi perilaku dan pengaturan eksperimental.

2.1.1 Asumsi perilaku lawan dalam *opponent modelling* untuk *Repeated Games*

Untuk mengoperasionalkan asumsi perilaku lawan secara konsisten pada pengaturan permainan yang beragam, tinjauan ini mengabstraksikan properti perilaku yang dapat diinferensi dari *interaction traces*. Secara khusus, perilaku lawan dikarakterisasi berdasarkan empat jenis dependensi yang dapat diamati: apakah aksi lawan (i) bervariasi lintas kondisi lingkungan dalam permainan yang sama, (ii) dimediasi oleh dinamika pada tingkat populasi, (iii) merespons secara langsung aksi agen, dan (iv) menunjukkan divergensi aksi pada riwayat interaksi terkini yang ekuivalen.

Kriteria-kriteria ini dievaluasi sepenuhnya pada tingkat perilaku eksternal dan tidak mengasumsikan adanya pengetahuan mengenai model internal lawan, mekanisme pembelajaran, maupun tujuan strategis yang dioptimalkan. Dengan demikian, kategorisasi yang digunakan tidak dimaksudkan sebagai taksonomi formal dari metode *opponent modelling*, melainkan sebagai kerangka analitis untuk memungkinkan

perbandingan lintas studi yang menggunakan paradigma pembelajaran, representasi strategi, dan abstraksi permainan yang berbeda.

Dalam studi yang tidak mendefinisikan asumsi perilaku lawan secara eksplisit, klasifikasi diturunkan secara konservatif berdasarkan pengaturan eksperimental dan dinamika interaksi yang dilaporkan. Tabel 2.2 merangkum hasil kategorisasi tersebut beserta referensi terkait.

Sejumlah pola konsisten muncul dari Tabel 2.2. Pertama, mayoritas studi terkini mengasumsikan lawan yang aksinya responsif terhadap agen dan menunjukkan divergensi perilaku pada riwayat interaksi yang setara Qiao et al. 2024; Lv et al. 2023; Li et al. 2025; Freire et al. 2023; Hu et al. 2023; Wang et al. 2019; De Weerd et al. 2022; Perera et al. 2025; Di et al. 2023. Pola ini mengindikasikan pergeseran fokus riset dari optimisasi terhadap strategi lawan yang tetap menuju ketahanan dan adaptasi terhadap lawan yang belajar atau berperilaku strategis secara dinamis.

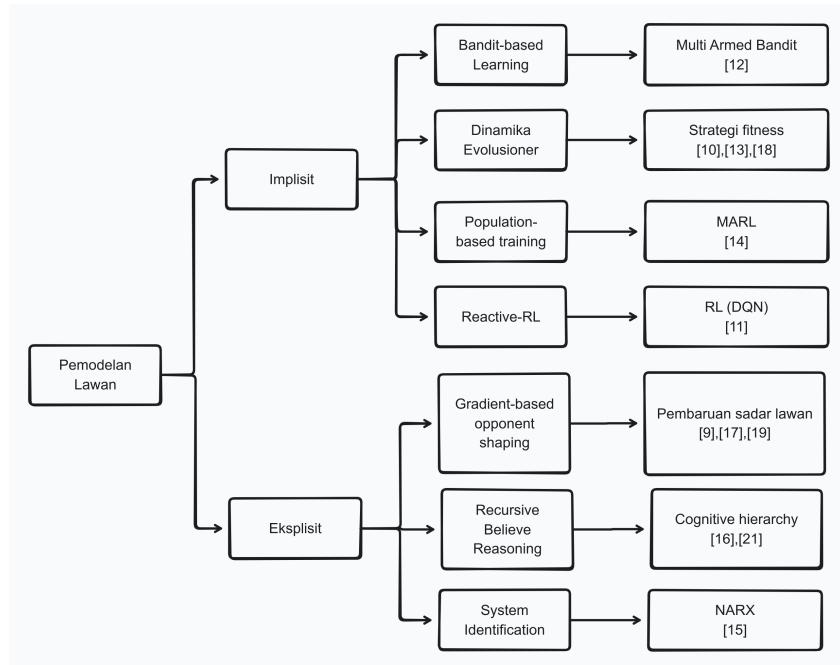
Kedua, dependensi perilaku yang dimediasi populasi terutama muncul dalam studi pada lingkungan evolusioner atau berbasis jaringan, di mana aksi individu dipengaruhi secara tidak langsung oleh dinamika agregat populasi Zhu et al. 2025; Gómez et al. 2025; Elhamer et al. 2020; Di et al. 2023; Perera et al. 2025. Dalam pengaturan ini, responsivitas terhadap agen sering kali terpisah dari perubahan strategi pada tingkat populasi, menghasilkan dinamika yang berbeda secara kualitatif dibandingkan skenario pembelajaran dua pemain.

Terakhir, hanya sebagian kecil karya yang secara simultan memodelkan perilaku lawan yang bergantung pada lingkungan, dimediasi populasi, dan responsif terhadap aksi agen Di et al. 2023. Keterbatasan ini menunjukkan bahwa lawan strategis yang sepenuhnya terkondisi oleh konteks interaksi—baik pada tingkat individu maupun kolektif—masih relatif kurang dieksplorasi, terutama dalam pengaturan *repeated games* dengan riwayat interaksi yang panjang dan tidak stasioner.

2.1.2 Pendekatan metodologis dalam *opponent modelling* untuk *Repeated Games*

Distribusi asumsi perilaku lawan pada Tabel 2.2 mencerminkan pergeseran metodologis yang lebih luas dalam riset *opponent modelling*, dari pengaturan dengan asumsi lawan yang tetap dan terkontrol menuju lawan yang adaptif, heterogen, dan berperilaku tidak stasioner. Namun, karakterisasi berbasis asumsi perilaku semata belum menjelaskan bagaimana kompleksitas tersebut dihadapi secara komputasional

di sisi agen. Untuk itu, Tabel ?? mereorganisasi studi-studi terdahulu berdasarkan paradigma pemodelan dominan dan mekanisme pembelajaran yang digunakan oleh agen.



Gambar 2.1 Diagram metodologi.

Pendekatan-pendekatan tersebut dapat dibedakan lebih lanjut berdasarkan apakah adaptasi terhadap lawan ditangani secara eksplisit atau implisit. Paradigma *opponent modelling* yang eksplisit—seperti *gradient-based opponent shaping* Qiao et al. 2024; Hu et al. 2023; Wang et al. 2019, *recursive belief reasoning* Freire et al. 2023; De Weerd et al. 2022, serta *system identification* Li et al. 2025—secara langsung membangun representasi internal perilaku lawan untuk memprediksi atau memengaruhi respons lawan di masa depan. Sebaliknya, pendekatan implisit, termasuk *reactive reinforcement learning* Lv et al. 2023, *population-based training* Perera et al. 2025, dinamika evolusioner Zhu et al. 2025; Di et al. 2023; Elhamer et al. 2020, dan *bandit-based learning-in-games* Gómez et al. 2025, beradaptasi terhadap lawan tanpa mempertahankan model perilaku lawan yang terpisah.

Meskipun pendekatan berbasis *reinforcement learning* dan *policy-gradient* mendominasi literatur terkini, paradigma tersebut umumnya menggabungkan pemodelan lawan ke dalam proses optimisasi kebijakan agen. Akibatnya, dinamika perilaku lawan sering kali terenkapsulasi secara implisit dalam parameter kebijakan atau

fungsi nilai, sehingga sulit untuk mengisolasi, menginterpretasi, atau memanfaatkan prediksi eksplisit terhadap respons lawan, terutama pada horizon interaksi yang panjang dan tidak stasioner.

Dalam konteks ini, pendekatan *system identification*, model sekuensial berbasis memori (misalnya LSTM) yang berfungsi sebagai pendekatan non-linear auto-regressive dengan input eksogen menawarkan alternatif metodologis yang berbeda. Alih-alih mengasumsikan struktur pembelajaran atau tujuan optimisasi lawan, LSTM memodelkan perilaku lawan sebagai proses dinamis yang dapat diinferensi langsung dari riwayat interaksi. Dengan memanfaatkan dependensi temporal dan aksi agen sebagai sinyal eksogen, pendekatan ini secara alami selaras dengan asumsi perilaku lawan yang responsif dan menunjukkan divergensi aksi pada riwayat interaksi yang ekuivalen, sebagaimana diidentifikasi pada Tabel 2.2.

Selain itu, LSTM memungkinkan pemisahan yang jelas antara proses prediksi perilaku lawan dan mekanisme pengambilan keputusan agen. Pemisahan ini memberikan fleksibilitas metodologis untuk menganalisis kualitas prediksi lawan secara independen dari kebijakan agen, serta memungkinkan integrasi dengan berbagai skema pengambilan keputusan tanpa memerlukan pelatihan ulang berbasis interaksi penuh seperti pada *reinforcement learning*. Karakteristik ini menjadikan pendekatan berbasis LSTM secara khusus menarik pada pengaturan dengan keterbatasan data, sumber daya komputasi, atau horizon waktu penelitian, sekaligus tetap mempertahankan kemampuan untuk menangkap dinamika perilaku lawan yang tidak stasioner dalam *repeated games*.

Sebagian besar pendekatan eksplisit sekilipun tetap melatih representasi lawan secara terintegrasi dengan objective reward agen. Literatur yang sepenuhnya memisahkan pembelajaran model lawan dari sinyal reward kebijakan relatif jarang dilaporkan, khususnya dalam konteks *repeated games* dengan lawan adaptif dan non-stasioner.

2.1.3 Bagaimana efektivitas strategi *opponent modelling* dievaluasi dalam *Repeated Games*?

Praktik evaluasi secara implisit mendefinisikan apa yang dianggap sebagai keberhasilan dalam interaksi multi-agent yang adaptif, baik dalam bentuk hasil kerja sama, ketahanan terhadap eksloitasi, stabilitas perilaku, maupun akurasi prediksi. Tabel 2.3 merangkum lingkungan evaluasi dan metrik yang digunakan dalam penelitian-

penelitian terdahulu, dengan tujuan mengidentifikasi pola evaluasi yang berulang serta aspek-aspek yang relatif terabaikan, alih-alih menetapkan standar normatif atau pemeringkatan kinerja.

Seperti terlihat pada Tabel 2.3, sebagian besar studi mengevaluasi efektivitas *opponent modelling* melalui metrik kinerja agregat jangka panjang, khususnya tingkat kerja sama Di et al. 2023; Elhamer et al. 2020; Wang et al. 2019; Jin et al. 2025, payoff rata-rata Lv et al. 2023; Perera et al. 2025; Li et al. 2025; Wang et al. 2019; De Weerd et al. 2022, serta konvergensi menuju equilibrium atau solusi stabil Gómez et al. 2025; Hu et al. 2023; Jin et al. 2025. Metrik-metrik ini secara inheren mengasumsikan interaksi berulang dengan horizon panjang, di mana kerugian eksplorasi pada tahap awal dapat dikompensasikan oleh perbaikan kinerja pada fase selanjutnya.

Namun demikian, asumsi horizon panjang ini membatasi daya representasi evaluasi terhadap skenario di mana interaksi bersifat terbatas, biaya eksplorasi signifikan, atau kesalahan awal sulit dipulihkan. Bahkan dalam studi yang mempertimbangkan lawan adaptif atau tidak stasioner, evaluasi umumnya dilakukan setelah fase pembelajaran mencapai stabilitas atau konvergensi Qiao et al. 2024, sehingga kinerja selama fase identifikasi lawan secara *online* relatif kurang diperhatikan.

Keterbatasan ini menjadi semakin relevan dalam pengaturan *repeated games* dengan horizon tetap yang pendek atau tidak pasti, di mana agen tidak dapat mengandalkan eksplorasi agresif tanpa risiko penurunan kinerja yang substansial. Dalam konteks tersebut, strategi eksplorasi yang terlalu invasif dapat menyebabkan salah koordinasi permanen, eksplorasi oleh lawan, atau kegagalan mencapai kerja sama sebelum interaksi berakhir. Oleh karena itu, evaluasi berbasis horizon panjang cenderung melebihkan keuntungan metode yang mengandalkan eksplorasi mendalam, sementara meremehkan pendekatan yang menekankan kehati-hatian dan efisiensi identifikasi perilaku lawan.

Sebagai respons terhadap celah ini, penggunaan horizon tetap yang pendek atau horizon stokastik dapat dipandang sebagai pilihan evaluasi yang lebih konservatif dan informatif. Horizon semacam ini secara eksplisit membatasi anggaran eksplorasi dan memaksa agen untuk menyeimbangkan antara identifikasi perilaku lawan dan kinerja langsung sejak tahap awal interaksi. Selain itu, horizon stokastik mengurangi insentif bagi strategi yang bergantung pada eksplorasi fase akhir permainan, sehingga mendorong perilaku yang lebih stabil dan berorientasi jangka pendek.

Dalam konteks *opponent modelling* berbasis prediksi eksplisit, evaluasi dengan horizon terbatas juga memungkinkan analisis yang lebih tajam terhadap kegu-

naan prediksi perilaku lawan. Alih-alih menilai keberhasilan hanya berdasarkan hasil agregat jangka panjang, pengaturan ini menyoroti seberapa cepat dan seberapa akurat model lawan dapat memberikan informasi yang berguna untuk pengambilan keputusan, serta sejauh mana agen mampu memanfaatkan prediksi tersebut tanpa melakukan eksplorasi yang berlebihan. Dengan demikian, praktik evaluasi ini memberikan perspektif pelengkap terhadap literatur yang ada, khususnya dalam menilai efisiensi dan kehati-hatian strategi *opponent modelling* pada pengaturan interaksi yang terbatas.

2.2 Kesimpulan

Bab ini telah meninjau literatur *opponent modelling* dalam interaksi strategis berulang dengan menelaah tiga dimensi utama, yaitu asumsi perilaku lawan, paradigma pemodelan yang digunakan, serta praktik evaluasi yang mendasari klaim keberhasilan. Fokus tinjauan diarahkan pada pengaturan *repeated games*, khususnya dilema sosial, di mana ketergantungan terhadap riwayat interaksi dan adaptasi perilaku menjadi aspek sentral dalam dinamika strategi (Axelrod dan Hamilton 1981).

Dari sisi asumsi perilaku, literatur menunjukkan kecenderungan untuk memodelkan lawan sebagai entitas yang adaptif dan responsif terhadap aksi agen (Hernandez-Leal et al. 2019). Namun, asumsi tersebut sering kali tertanam secara implisit dalam desain algoritma atau skema pelatihan, tanpa pemisahan yang jelas antara dinamika perilaku lawan dan mekanisme optimisasi kebijakan agen. Kondisi ini menyulitkan analisis terpisah terhadap kualitas representasi internal perilaku lawan.

Secara metodologis, pendekatan *opponent modelling* dalam MARL dapat dikategorikan ke dalam paradigma implisit dan eksplisit. Pendekatan implisit umumnya mengintegrasikan dinamika lawan langsung ke dalam parameter kebijakan melalui pelatihan berbasis reward atau *policy gradient* (Hernandez-Leal et al. 2019). Semen-tara itu, pendekatan eksplisit mempertahankan representasi khusus mengenai perilaku lawan, namun dalam banyak kasus tetap dilatih secara terintegrasi dengan objective reward agen (Albrecht dan Stone 2018). Akibatnya, kualitas prediksi perilaku lawan sulit dievaluasi secara independen dari kinerja kebijakan akhir.

Selain itu, sebagian besar model lawan berfokus pada prediksi satu langkah ke depan (*one-step prediction*) (Hernandez-Leal et al. 2019), dengan penekanan pada stabilitas jangka panjang atau konvergensi menuju perilaku tertentu. Pendekatan ini kurang mengeksplorasi pemodelan multi-langkah dalam horizon pendek, padahal dalam pengaturan interaksi terbatas atau tidak pasti, dinamika jangka pendek memiliki

pengaruh signifikan terhadap kinerja keseluruhan.

Dari sisi evaluasi, praktik yang dominan masih mengandalkan metrik agregat jangka panjang seperti tingkat kerja sama, payoff rata-rata, dan konvergensi menuju equilibrium. Evaluasi semacam ini secara implisit mengasumsikan bahwa biaya eksplorasi awal dapat dikompensasi dalam fase interaksi berikutnya. Namun, dalam pengaturan dengan horizon tetap yang pendek atau terminasi stokastik, eksplorasi yang tidak terarah dapat menimbulkan penurunan kinerja yang sulit dipulihkan (Nisan et al. 2008). Dengan demikian, pendekatan yang mengandalkan eksplorasi agresif berpotensi terlihat unggul dalam evaluasi horizon panjang, tetapi kurang sesuai untuk skenario interaksi terbatas.

Lebih lanjut, meskipun beberapa studi mempertimbangkan ketidakpastian atau variasi perilaku lawan, pemanfaatan ukuran ketidakpastian prediktif secara eksplisit dalam mekanisme seleksi aksi masih relatif terbatas. Ketidakpastian sering kali muncul sebagai efek samping dari proses pembelajaran, bukan sebagai komponen yang secara langsung dipertimbangkan dalam perhitungan skor aksi sebelum keputusan diambil.

Berdasarkan sintesis tersebut, dapat diidentifikasi celah penelitian pada pengembangan pendekatan *opponent modelling* yang: (i) memisahkan pelatihan model lawan dari sinyal reward kebijakan, (ii) melakukan prediksi multi-langkah dalam horizon pendek, dan (iii) memanfaatkan ukuran ketidakpastian prediktif secara eksplisit dalam proses seleksi aksi, khususnya pada pengaturan interaksi berulang dengan horizon terbatas atau terminasi stokastik.

Secara khusus, literatur masih sangat terbatas dalam membahas pengaturan *repeated games* dengan terminasi stokastik (misalnya distribusi geometrik), di mana horizon interaksi bersifat tidak pasti dan eksplorasi harus dilakukan secara hati-hati sejak awal interaksi.

Celah ini membuka ruang bagi pendekatan yang menekankan pemodelan perilaku lawan berbasis observasi historis secara terpisah dari optimisasi kebijakan, serta evaluasi yang sensitif terhadap trade-off antara kinerja langsung dan pengurangan ketidakpastian prediksi dalam fase awal interaksi.

Tabel 2.1 Ringkasan Penelitian Terkait

Ref	Nama Model	Temuan Utama
Qiao et al. 2024	Online Opponent Modeling (O2M)	O2M mampu beradaptasi lebih baik dan memperoleh rata-rata reward yang lebih tinggi dibandingkan model baseline.
2	Zero Determinant Strategy under Evolutionary Dynamic	Proporsi akhir kerja sama melebihi strategi Extort, menunjukkan keunggulan dalam dinamika evolusioner.
3	Hierarchical Gifting DQN	Mampu mendeteksi perubahan strategi lawan secara real-time dan secara dinamis menyesuaikan incentif kerja sama.
4	Teamwork Game + MAMAB	Agregasi NE teoretis ($\chi^2 = 0.992$); mereproduksi pola mirip manusia (social loafing, compensation);
5	Environment–Strategy Coupling Model	Kopling antara umpan balik lingkungan dan dinamika strategi secara signifikan meningkatkan stabilitas kerja sama kelompok;
6	Ensemble-Training Cooperative Agent	PRobust dan memiliki generalisasi lebih baik dibanding pelatihan terhadap lawan tunggal; namun dengan biaya komputasi lebih tinggi dan potensi estimasi reward intrinsik yang sub-optimal.
7	LSTM-Strategy	RNN mampu mengaproksimasi dinamika no-regret yang halus secara akurat serta memungkinkan eksplorasi yang menguntungkan pada tingkat non-stasioneritas rendah.
8	Internal Model	konvergensi lebih cepat dan stabil dibanding TD-learning standar; performa menuju ketika perilaku lawan bergantung pada aksi agen sendiri (misalnya Tit-for-Tat).
9	Symmetric Learning Awareness (SLA)	Keseimbangan yang lebih stabil dan menghindari perilaku siklik yang muncul pada metode gradien standar; pemodelan eksplisit meningkatkan konvergensi dan stabilitas.
10	Extended Social Particle Swarm (SPS) Model	Tingkat pembaruan informasi yang tinggi menghasilkan klaster kerja sama yang lebih dinamis dan beragam; Ukuran populasi dan laju pembaruan informasi secara bersama-sama membentuk pola kerja sama.
11	Deep Multiagent Reinforcement Learning (untuk SPD)	Mencapai kerja sama mutual dalam self-play; menghindari eksplorasi oleh lawan defektif; lebih adaptif terhadap perubahan strategi lawan.
12	Suggestion Sharing (SS)	Mencapai tingkat kerja sama kompetitif atau lebih baik dibanding value sharing, policy sharing, dan intrinsic reward; menjadi alternatif untuk mendukung kerja sama.

Tabel 2.2 Asumsi perilaku lawan berdasarkan dependensi perilaku yang dapat diamati.

Env.	Pop.	Agent	Div.	Kategori Perilaku Lawan	Ref.
—	—	✓	—	Reactive	Jin et al. 2025
—	✓	—	—	Population-Conformist	Gómez et al. 2025
—	✓	✓	—	Contextual Reactive	Elhamer et al. 2020
—	—	✓	✓	Learning Opponent	Qiao et al. 2024; Lv et al. 2023; Li et al. 2025; Freire et al. 2023; Hu et al. 2023; Wang et al. 2019; De Weerd et al. 2022
—	✓	✓	✓	Population-Contextual Strategic	Perera et al. 2025
✓	✓	—	✓	Heterogeneous Collective Behavior	Zhu et al. 2025
✓	✓	✓	✓	Environment-Conditioned Strategic	Di et al. 2023

Catatan:

Env. — perilaku bervariasi lintas lingkungan dalam permainan yang sama;

Pop. — perilaku dimediasi oleh interaksi tingkat populasi;

Agent — perilaku merespons secara langsung aksi agen;

Div. — divergensi aksi terjadi pada riwayat interaksi terkini yang ekuivalen.

Tanda centang menunjukkan adanya dependensi.

Tabel 2.3 Lingkungan evaluasi dan metrik dalam penelitian

Ref.	Lingkungan Evaluasi	Metrik
Qiao et al. 2024	Self-play simetris	MSE selama pelatihan offline; akurasi memori laten
Zhu et al. 2025	Jaringan scale-free dengan strategi zero-determinant	Frekuensi kerja sama (C) dan eksplorasi (E)
Lv et al. 2023	Opponent adaptif (dirata-ratakan pada beberapa opponent)	Nilai reward
Gómez et al. 2025	Simulator teamwork-game khusus (aggregative public good games); eksperimen sintetis	Produktivitas tim agregat; uji kecocokan χ^2 terhadap equilibrium; konvergensi ke Nash equilibrium; kontribusi individu
Di et al. 2023	Simulator evolutionary game pada jaringan terstruktur	Tingkat kerja sama; fraksi kooperator; ambang fase transisi
Perera et al. 2025	Repeated matrix games dengan populasi opponent sintetis	Payoff rata-rata; tingkat kerja sama; robustness terhadap himpunan opponent; generalisasi
Li et al. 2025	Repeated zero-sum games melawan Hedge, OMD, dan Regret Matching	Galat prediksi; payoff kumulatif; robustness terhadap non-stationarity
Freire et al. 2023	Repeated matrix games; simulasi robotik embodied waktu-kontinu	Efektivitas; stabilitas; akurasi prediksi
Hu et al. 2023	Simulasi repeated matrix game	Payoff rata-rata; kecepatan konvergensi; pemilihan equilibrium
Elhamer et al. 2020	Simulasi continuous-space skala besar (FLAME GPU)	Tingkat kerja sama; ukuran dan jumlah klaster kooperatif; kecepatan agen; stabilitas klaster
Wang et al. 2019	Lingkungan SPD 2D khusus (Fruit Gathering; Apple—Pear games)	Reward individu rata-rata; total kesejahteraan sosial; akurasi deteksi derajat kerja sama
Jin et al. 2025	Benchmark MARL (Cleanup, Harvest, Sequential PD, Tragedy of the Commons)	Return ternormalisasi; tingkat kerja sama; kecepatan konvergensi; perbedaan kebijakan (MSE)
De Weerd et al. 2022	Simulasi Colored Trails dengan peningkatan ketidakpastian lingkungan	Skor allocator; skor responder; total kesejahteraan sosial

BAB III

LANDASAN TEORI

3.1 Dilema Sosial

Dilema sosial merupakan situasi di mana keputusan rasional secara individual menghasilkan luaran yang tidak optimal secara kolektif(Axelrod dan Hamilton 1981). Secara formal, kondisi ini dapat dinyatakan sebagai:

$$\sum_{i=1}^n u_i(a_i, a_{-i}) < \sum_{i=1}^n u_i(a'_i, a'_{-i}) \quad (3.1)$$

di mana n adalah jumlah pemain, a_i adalah aksi rasional individu dan a'_i adalah profil aksi yang memaksimalkan kesejahteraan kolektif.

Formulasi ini menjelaskan adanya konflik antara rasionalitas individu dan optimalitas sosial. Namun, ekspresi agregat tersebut belum menyediakan struktur analitis yang cukup untuk memodelkan interaksi strategis antar agen secara eksplisit.

3.2 Game Theory

teori permainan memberikan kerangka formal yang memodelkan pemain, strategi, dan payoff secara terstruktur Bonanno 2024, dalam bentuk normal didefinisikan sebagai:

$$G = (N, A, U) \quad (3.2)$$

dengan N himpunan pemain, $A = A_1 \times A_2$ himpunan profil strategi atau *Action space*, dan $U = (u_1, u_2)$ fungsi payoff atau *utility*.

Nash Equilibrium adalah profil strategi a^* yang memenuhi:

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*) \quad \forall a_i \in A_i \quad (3.3)$$

Menunjukkan pilihan aksi selain a_i^* tidak dapat memberikan keuntungan lebih besar dari strategi keseimbangan. Kerangka ini memungkinkan analisis rasionalitas strategis dalam interaksi statik. Namun, model bentuk normal bersifat satu tahap dan mengasumsikan struktur payoff tetap.

3.3 Prisoner's Dilemma

Prisoner's Dilemma merepresentasikan konflik rasionalitas individu dan kollektif secara eksplisit Axelrod dan Hamilton 1981. Aksi yang dapat dipilih adalah *Cooperate (C)* atau *Defect (D)*. Struktur payoff Prisoner's Dilemma diberikan oleh:

	<i>C</i>	<i>D</i>	
<i>C</i>	(R, R)	(S, T)	
<i>D</i>	(T, S)	(P, P)	

(3.4)

Dengan *R* (Reward) adalah payoff jika kedua pemain bekerja sama, *T* (Temptation) adalah payoff bagi pemain yang berkhianat sementara yang lain bekerja sama, *S* (Sucker's payoff) adalah payoff bagi pemain yang bekerja sama sementara yang lain berkhianat, dan *P* (Punishment) adalah payoff jika kedua pemain berkhianat. dengan ketidaksamaan:

$$T > R > P > S, \quad 2R > T + S \quad (3.5)$$

Dalam permainan satu tahap, strategi dominan adalah defeksi (*D*), sehingga keseimbangan Nash berada pada (*D*, *D*) meskipun (*C*, *C*) lebih optimal secara kollektif.

Namun, interaksi nyata jarang terjadi hanya satu kali.

3.4 Iterated Prisoner's Dilemma

Permainan PD diperluas menjadi bentuk berulang disebut Iterated Prisoner's Dilemma (IPD), permainan diulang selama *T* atau *Turn* tahap dengan payoff kumulatif atau terdiskonto. Dalam IPD, strategi dapat bergantung pada histori interaksi, memungkinkan untuk pembalasan dan kerja sama yang berkelanjutan.

3.4.1 Finite Horizon

Pada formulasi finite horizon dengan panjang permainan tetap *T*, utilitas total pemain didefinisikan sebagai penjumlahan utilitas pada setiap ronde:

$$U_i = \sum_{t=1}^T u_i(a_i^t, a_{-i}^t) \quad (3.6)$$

Pada model ini, tidak digunakan faktor diskonto. Namun, secara teoretis, pendekatan ini cenderung menghasilkan strategi defeksi melalui mekanisme *backward induction*.

3.4.2 Infinite Horizon dan Discount Factor

Alternatif formulasi adalah infinite horizon, di mana permainan berlangsung tanpa batas dengan faktor diskonto $\delta \in (0, 1]$. Utilitas pemain didefinisikan sebagai:

$$U_i = \sum_{t=1}^{\infty} \delta^{t-1} u_i(a_i^t, a_{-i}^t) \quad (3.7)$$

Faktor diskonto δ merepresentasikan preferensi terhadap reward di masa depan, di mana nilai yang lebih kecil menunjukkan orientasi jangka pendek, sedangkan nilai mendekati 1 menunjukkan orientasi jangka panjang.

IPD memungkinkan strategi adaptif berbasis riwayat (Axelrod dan Hamilton 1981). Namun, analisis keseimbangan klasik tetap mengasumsikan strategi tetap dan rasionalitas sempurna.

3.4.3 Stochastic Termination

Dalam banyak interaksi nyata, panjang permainan tidak tetap, melainkan mengikuti mekanisme terminasi stokastik. Misalkan permainan berlanjut dari tahap t ke $t + 1$ dengan probabilitas tetap $\gamma \in (0, 1)$:

$$P(\text{continue at } t+1 \mid t) = \gamma \quad (3.8)$$

Maka panjang permainan mengikuti distribusi geometrik, dan ekspektasi payoff menjadi:

$$\mathbb{E}[U_i] = \sum_{t=1}^{\infty} \gamma^{t-1} u_i(a_i^t, a_{-i}^t) \quad (3.9)$$

Formulasi ini ekuivalen dengan permainan berulang dengan faktor diskonto γ (Sutton dan Barto 2015), namun memiliki interpretasi probabilistik sebagai proses berhenti geometrik. Ketidakpastian horizon ini mempengaruhi nilai eksplorasi sejak tahap awal interaksi.

3.4.4 Hubungan Stochastic Termination dan Discounted Return

Dalam repeated game dengan probabilitas terminasi tetap $1 - \gamma$, interaksi berlanjut ke periode berikutnya dengan probabilitas γ (Sutton dan Barto 2015). Struktur ini ekuivalen secara matematis dengan discounted infinite-horizon return:

$$Q(a_t) = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k u_i(a_i^{t+k}, a_{-i}^{t+k}) \right] \quad (3.10)$$

Dengan demikian, faktor diskonto γ dapat diinterpretasikan sebagai probabilitas kelanjutan permainan, sehingga discounted return merepresentasikan ekspektasi utilitas dalam repeated game dengan horizon stokastik.

Penjelasan ini menghubungkan struktur teoretis permainan berulang dengan formulasi objektif evaluasi nilai yang digunakan dalam penelitian ini.

3.4.5 Definisi History Interaksi pada mekanisme online

Dalam konteks iterated game dua pemain, history hingga waktu t tidak hanya terdiri dari aksi satu pemain, tetapi pasangan aksi kedua pemain pada setiap putaran.

Secara formal, history didefinisikan sebagai:

$$h_t = ((a_i^1, a_{-i}^1), (a_i^2, a_{-i}^2), \dots, (a_i^t, a_{-i}^t)) \quad (3.11)$$

dengan a_i^k menyatakan aksi pemain i pada waktu k , dan a_{-i}^k menyatakan aksi lawan pada waktu yang sama.

Dalam pengaturan iteratif, history tidak tersedia secara sekaligus, melainkan terbentuk secara *online*. Pada setiap putaran t , history diperbarui secara inkremental sebagai:

$$h_t = (h_{t-1}, (a_i^t, a_{-i}^t)) \quad (3.12)$$

dengan $h_0 = \emptyset$.

Formulasi ini menegaskan bahwa agen tidak memiliki akses terhadap trajectory interaksi di masa depan, dan hanya dapat menggunakan informasi yang telah terobservasi hingga waktu berjalan. Dengan kata lain, proses pengambilan keputusan berlangsung dalam kerangka kausal dan sekuensial.

History interaksi menyediakan rekaman lengkap atas realisasi aksi kedua pemain. *Namun*, history hanya merepresentasikan keluaran observabel dari proses pe-

ngambilan keputusan lawan, bukan mekanisme generatif yang mendasarinya. Agen tidak memiliki akses langsung terhadap strategi internal, parameter keputusan, atau pun aturan adaptasi yang digunakan oleh lawan.

Dengan demikian, meskipun seluruh pasangan aksi teramat secara bertahap, struktur strategi lawan tetap bersifat laten. Kondisi ini menyebabkan interaksi dalam IPD tidak memenuhi asumsi informasi lengkap sebagaimana pada analisis keseimbangan statik klasik.

Oleh karena itu, permasalahan strategis dalam IPD dengan agen adaptif berada dalam kerangka permainan dengan informasi tidak lengkap, di mana agen harus melakukan inferensi terhadap strategi lawan secara bertahap seiring pertambahan history.

3.4.6 Incomplete Information dan Pembentukan Belief

Dalam permainan dengan informasi tidak lengkap, ketidakpastian terhadap strategi lawan direpresentasikan melalui distribusi probabilitas atas kemungkinan tipe atau parameter strategi. Alih-alih mengasumsikan strategi tetap dan diketahui, agen memelihara suatu *belief state* yang diperbarui seiring bertambahnya histori interaksi.

Secara konseptual, belief state pada waktu t dapat dituliskan sebagai:

$$b_t = P(\theta \mid h_t) \quad (3.13)$$

dengan θ merepresentasikan representasi laten dari strategi lawan, dan h_t adalah histori interaksi hingga waktu t .

Formulasi ini menegaskan bahwa pengambilan keputusan dalam IPD adaptif bukan sekadar persoalan memilih aksi optimal terhadap strategi tetap, melainkan proses pembaruan belief secara sekuensial terhadap dinamika perilaku lawan. Dengan demikian, fokus analisis bergeser dari pencarian equilibrium statik menuju inferensi dinamis berbasis histori.

3.5 Pemodelan Lawan

Pemodelan lawan atau *opponent modelling* memungkinkan untuk mempelajari perilaku ataupun strategi lawan (Shoham 2009).

Diberikan riwayat interaksi penuh:

$$h_t = ((a_i^1, a_{-i}^1), \dots, (a_i^t, a_{-i}^t)) \quad (3.14)$$

model parametrik f_θ mempelajari distribusi kondisional:

$$p_\theta(a_{t+1}^{-i} \mid h_t) \quad (3.15)$$

Pendekatan regresi linear sederhana mampu memodelkan hubungan statik, namun memiliki keterbatasan dalam menangkap dependensi temporal dan pola nonlinier.

3.6 RNN

Arsitektur berbasis jaringan saraf dapat menangkap dependensi temporal dan nonlinier untuk memodelkan dinamika yang lebih kompleks (Hochreiter dan Schmidhuber 1997). RNN memperbarui keadaan tersembunyi sebagai:

3.6.1 Recurrent Neural Network untuk Opponent Modelling

RNN digunakan untuk memodelkan dinamika perilaku lawan berdasarkan urutan observasi. State tersembunyi diperbarui pada setiap waktu sebagai berikut:

$$h_t = \phi(Wx_t + Uh_{t-1} + b) \quad (3.16)$$

di mana:

- x_t adalah input pada waktu t , yang merepresentasikan observasi interaksi (misalnya aksi pemain dan lawan pada ronde sebelumnya),
- h_t adalah hidden state yang merepresentasikan belief terhadap strategi lawan hingga waktu t ,
- h_{t-1} adalah hidden state pada waktu sebelumnya,
- W adalah matriks bobot untuk input,
- U adalah matriks bobot rekuren yang menghubungkan state sebelumnya,
- b adalah bias,
- ϕ adalah fungsi aktivasi non-linear.

3.6.2 Latent Belief Representation

Dalam konteks opponent modelling, state tersembunyi h_t pada RNN diinterpretasikan sebagai representasi laten dari belief terhadap strategi lawan. Secara formal, belief terhadap aksi lawan pada waktu t dapat dimodelkan sebagai distribusi probabilitas bersyarat:

$$b_t(a_{-i}) = P(a_{-i}^t \mid h_t) \quad (3.17)$$

di mana h_t merupakan representasi latent yang merangkum seluruh histori interaksi hingga waktu t . Untuk memperoleh estimasi distribusi aksi lawan, digunakan fungsi pemetaan sebagai berikut:

$$\hat{b}_t = \text{softmax}(Vh_t) \quad (3.18)$$

di mana:

- $b_t(a_{-i})$ adalah belief terhadap aksi lawan,
- h_t adalah latent belief representation,
- V adalah matriks bobot output,
- \hat{b}_t adalah estimasi distribusi probabilitas aksi lawan.

Dengan demikian, h_t tidak secara eksplisit merepresentasikan strategi lawan, melainkan embedding latent yang digunakan untuk mengaproksimasi belief tersebut dalam pemodelan lawan. Namun, pada RNN dengan dependensi jangka panjang menimbulkan masalah *vanishing gradient* yaitu gradient yang menumpuk panjang dan menggecil membuat model tidak dapat belajar dependensi jangka panjang dengan baik.

3.7 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) merupakan pengembangan dari Recurrent Neural Network (RNN) yang dirancang untuk menangkap dependensi jangka panjang melalui mekanisme memori eksplisit. Dalam konteks *opponent modelling*, LSTM digunakan untuk membangun representasi latent terhadap strategi lawan berdasarkan histori interaksi.

Diberikan input x_t (observasi pada waktu t , misalnya aksi lawan), hidden state sebelumnya h_{t-1} (representasi laten sebelumnya), dan parameter bobot W , U , serta bias b , setiap komponen LSTM bekerja secara berurutan sebagai berikut.

3.7.1 Input Gate

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3.19)$$

Input $x_t \in \mathbb{R}^d$ merepresentasikan observasi saat ini, sedangkan $h_{t-1} \in \mathbb{R}^h$ adalah ringkasan historis interaksi sebelumnya. Namun, tidak semua informasi baru relevan atau stabil untuk memperbarui belief terhadap strategi lawan. Oleh karena itu, input gate $i_t \in [0, 1]^h$ mengontrol seberapa besar setiap dimensi informasi baru akan diterima ke dalam memori, dengan $\sigma(\cdot)$ sebagai fungsi sigmoid dan W_i, U_i, b_i sebagai parameter yang dipelajari.

3.7.2 Candidate Cell State

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (3.20)$$

Informasi mentah dari x_t tidak langsung disimpan karena dapat mengandung noise atau pola sementara yang belum representatif. Namun, model tetap membutuhkan representasi kandidat dari informasi baru tersebut. Oleh karena itu, $\tilde{c}_t \in [-1, 1]^h$ dibentuk melalui transformasi non-linear $\tanh(\cdot)$ sebagai kandidat memori baru, dengan parameter W_c, U_c, b_c .

3.7.3 Forget Gate

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (3.21)$$

Cell state sebelumnya c_{t-1} menyimpan histori panjang interaksi yang membentuk belief terhadap lawan. Namun, tidak semua informasi lama tetap relevan karena strategi lawan dapat berubah. Oleh karena itu, forget gate $f_t \in [0, 1]^h$ menentukan bagian mana dari memori lama yang perlu dipertahankan atau dilupakan secara adaptif.

3.7.4 Cell State (Memori Jangka Panjang)

Cell state $c_t \in \mathbb{R}^h$ merupakan jalur utama propagasi informasi jangka panjang dalam LSTM yang berfungsi sebagai *latent belief representation* terhadap strategi lawan. Namun, memori ini harus mampu menjaga stabilitas sekaligus tetap adaptif terhadap informasi baru. Oleh karena itu, pembaruannya dirumuskan sebagai:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (3.22)$$

Operator \odot menyatakan perkalian elemen-per-elemen. Komponen $f_t \odot c_{t-1}$ mempertahankan informasi lama yang masih relevan, sedangkan $i_t \odot \tilde{c}_t$ menambahkan informasi baru secara selektif. Struktur aditif ini penting karena menjaga stabilitas gradien dan memungkinkan pembelajaran dependensi jangka panjang.

3.7.5 Output Gate

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (3.23)$$

Meskipun c_t menyimpan informasi lengkap, tidak seluruhnya relevan untuk prediksi saat ini. Namun, model perlu mengekstrak bagian informasi yang paling informatif untuk pengambilan keputusan. Oleh karena itu, output gate $o_t \in [0, 1]^h$ mengontrol seberapa besar informasi dari cell state akan diekspos ke hidden state, dengan parameter W_o, U_o, b_o .

3.7.6 Hidden State Update

$$h_t = o_t \odot \tanh(c_t) \quad (3.24)$$

Cell state c_t mengandung representasi lengkap jangka panjang, namun terlalu kaya untuk digunakan secara langsung. Oleh karena itu, hidden state $h_t \in \mathbb{R}^h$ dibentuk sebagai representasi latent terfilter yang lebih terfokus, yang kemudian digunakan untuk memprediksi aksi lawan atau sebagai input ke modul pengambilan keputusan.

Struktur ini memungkinkan model mempertahankan dependensi jangka panjang sekaligus tetap adaptif terhadap perubahan strategi lawan. Oleh karena itu, LSTM sangat sesuai untuk digunakan dalam permainan berulang seperti Iterated Prisoner's Dilemma, di mana perilaku lawan bergantung pada histori interaksi. Namun,

karena LSTM hanya menghasilkan prediksi satu langkah ke depan, model ini belum secara eksplisit mempertimbangkan konsekuensi jangka panjang, sehingga masih berpotensi menghasilkan kebijakan yang suboptimal.

3.8 Prediksi Multi-Langkah

Model LSTM menghasilkan distribusi prediktif satu langkah ke depan $p(a_{t+1}^{-i} | h_t)$. Namun, dalam permainan berulang dengan terminasi stokastik, nilai suatu aksi tidak hanya ditentukan oleh respons lawan pada tahap berikutnya, melainkan oleh konsekuensi jangka panjang sepanjang lintasan interaksi.

3.8.1 Forecasting Rekursif dan Monte Carlo Rollout

Untuk memperoleh estimasi lintasan masa depan, model digunakan secara rekursif:

$$\hat{a}_{t+k}^{-i} \sim p_\theta(\cdot | \hat{h}_{t+k-1}) \quad (3.25)$$

Pendekatan ini menghasilkan distribusi atas lintasan aksi $\{\hat{a}_{t+1}^{-i}, \hat{a}_{t+2}^{-i}, \dots\}$. Karena ruang lintasan tumbuh secara eksponensial, ekspektasi payoff dihitung melalui simulasi Monte Carlo:

$$\hat{Q}(a_t) = \frac{1}{M} \sum_{m=1}^M \sum_{k=0}^{\infty} \gamma^k u_i^{(m)}(a_i^{t+k}, a_{-i}^{t+k}) \quad (3.26)$$

dengan M jumlah simulasi independen.

Dengan demikian, evaluasi nilai aksi diperoleh secara empiris melalui sampling lintasan interaksi.

3.8.2 Dari Akurasi Prediksi ke Kinerja Strategis

Namun demikian, akurasi prediksi distribusi aksi lawan tidak secara langsung menjamin kinerja strategis yang optimal. Tujuan agen dalam permainan berulang bukan sekadar meminimalkan kesalahan prediksi, melainkan memaksimalkan payoff kumulatif.

Dalam formulasi permainan, misalkan i menyatakan agen yang dikontrol, dan $-i$ menyatakan himpunan seluruh lawan. Aksi yang diambil agen pada waktu t dinotasikan sebagai $a_t^i \in A_i$, sedangkan aksi lawan dinotasikan sebagai $a_t^{-i} \in A_{-i}$. Fungsi

utilitas $u_i(a_t^i, a_t^{-i})$ merepresentasikan payoff yang diterima agen i ketika ia memilih aksi a_t^i dan lawan memilih aksi a_t^{-i} . Dalam konteks *opponent modelling*, a_t^{-i} merupakan variabel yang tidak dapat dikontrol dan hanya dapat diperkirakan melalui belief yang dibangun model.

Cumulative Regret (Classical Regret) Untuk mengukur performa jangka panjang, digunakan metrik regret kumulatif (Nisan et al. 2008):

$$R_T = \max_{a_i \in A_i} \sum_{t=1}^T u_i(a_i, a_t^{-i}) - \sum_{t=1}^T u_i(a_t^i, a_t^{-i}) \quad (3.27)$$

Pada persamaan di atas, $\max_{a_i \in A_i}$ merepresentasikan aksi tetap terbaik (*best fixed action*) yang dipilih secara retrospektif setelah seluruh interaksi hingga horison T diamati. Namun, aksi optimal ini tidak diketahui selama permainan berlangsung. Term pertama $\sum_{t=1}^T u_i(a_i, a_t^{-i})$ merepresentasikan total payoff yang akan diperoleh jika agen selalu memainkan aksi terbaik tersebut, sedangkan term kedua $\sum_{t=1}^T u_i(a_t^i, a_t^{-i})$ adalah payoff aktual dari aksi yang benar-benar diambil agen. Oleh karena itu, R_T mengukur kerugian akibat tidak mengetahui strategi optimal sejak awal.

Pada permainan berulang, optimalitas strategi tidak hanya ditentukan oleh kualitas aksi pada satu waktu, melainkan oleh konsistensi keputusan sepanjang interaksi. Namun, regret kumulatif membandingkan performa agen terhadap satu aksi tetap terbaik (*best fixed action*), yang dipilih secara retrospektif.

Pendekatan ini bersifat terbatas, karena pembanding berupa aksi statis tidak mampu merepresentasikan strategi kondisional yang bergantung pada histori interaksi. Dalam permainan seperti Iterated Prisoner's Dilemma, strategi optimal sering kali berbentuk kebijakan adaptif, di mana aksi pada waktu tertentu bergantung pada perilaku lawan sebelumnya.

Akibatnya, aksi yang optimal secara per-langkah tidak selalu menghasilkan payoff kumulatif yang optimal, karena keputusan saat ini dapat mempengaruhi respons lawan di masa depan. Dengan demikian, evaluasi berbasis aksi tetap menjadi kurang representatif terhadap kualitas strategi dalam konteks permainan berulang.

Equilibrium Regret Oleh karena itu, digunakan konsep equilibrium regret, yang mengukur deviasi performa agen terhadap strategi ekuitibrium yang mempertimbangkan interaksi strategis antar agen.

Secara umum, equilibrium regret dapat didefinisikan sebagai selisih antara payoff yang diperoleh agen dengan payoff yang akan diperoleh jika agen mengikuti strategi ekuilibrium. Misalkan π_i^* adalah strategi ekuilibrium untuk agen i , dan π_i adalah strategi yang digunakan, maka equilibrium regret dapat dituliskan sebagai:

$$R_T^{eq} = \sum_{t=1}^T (u_i(a_t^*, a_t^{-i}) - u_i(a_t^i, a_t^{-i})) \quad (3.28)$$

di mana a_t^* merupakan aksi yang dihasilkan oleh strategi ekuilibrium π_i^* pada waktu t .

Meskipun equilibrium regret memberikan pembanding yang lebih representatif dibandingkan aksi tetap, pendekatan ini tetap mengasumsikan keberadaan strategi ekuilibrium yang stabil sepanjang interaksi. Namun, dalam praktiknya, agen berinteraksi dengan lawan yang adaptif, di mana strategi lawan dapat berubah sebagai respons terhadap histori permainan dan tindakan agen itu sendiri. Dengan demikian, bahkan strategi ekuilibrium atau respons terbaik yang relevan dapat beralih dari waktu ke waktu.

Dynamic Regret Oleh karena itu, diperlukan metrik evaluasi yang mampu menangkap perubahan optimalitas secara temporal, tanpa mengasumsikan strategi pembanding yang statis. Untuk tujuan tersebut, digunakan dynamic regret yang membandingkan performa agen dengan aksi optimal yang dapat berubah pada setiap waktu. Oleh karena itu digunakan dynamic regret:

$$R_T^{dyn} = \sum_{t=1}^T (u_i(a_t^*, a_t^{-i}) - u_i(a_t^i, a_t^{-i})) \quad (3.29)$$

di mana $a_t^* \in A_i$ adalah aksi optimal pada waktu t yang didefinisikan sebagai:

$$a_t^* = \arg \max_{a_i \in A_i} u_i(a_i, a_t^{-i}) \quad (3.30)$$

Dengan demikian, dynamic regret membandingkan performa agen dengan strategi optimal yang dapat berubah di setiap waktu, sehingga lebih sesuai untuk mengevaluasi kinerja dalam lingkungan yang adaptif.

Meskipun nilai aksi secara teoritis sering didefinisikan dalam horizon tak hingga dengan faktor diskonto $\gamma \in (0, 1)$, evaluasi empiris dilakukan dalam horizon terbatas T yang merepresentasikan panjang simulasi interaksi. Pendekatan ini tetap

konsisten, karena untuk $\gamma < 1$, kontribusi payoff masa depan menurun secara eksponensial terhadap waktu, sehingga kontribusi pada horizon sangat panjang menjadi semakin kecil.

3.8.3 Decoupled Opponent Modelling

Dalam konteks *opponent modelling*, terdapat dua komponen utama, yaitu model prediksi perilaku lawan dan kebijakan (*policy*) agen. Secara umum, kedua komponen ini dapat dirancang secara terintegrasi maupun terpisah.

Pendekatan terintegrasi mempelajari representasi lawan secara end-to-end bersama dengan policy, sehingga prediksi yang dihasilkan langsung digunakan dalam proses pengambilan keputusan. Pendekatan ini banyak digunakan dalam *reinforcement learning*, karena memungkinkan optimasi langsung terhadap tujuan akhir berupa reward kumulatif.

Namun, keterkaitan yang erat antara prediksi dan policy menyulitkan interpretasi kualitas model lawan secara terpisah. Kesalahan prediksi dapat terkompensasi oleh policy, atau sebaliknya, sehingga evaluasi berbasis performa akhir tidak selalu mencerminkan akurasi representasi lawan.

Sebagai alternatif, pendekatan *decoupled opponent modelling* memisahkan proses pembelajaran model lawan dari kebijakan agen. Dalam kerangka ini, model bertujuan untuk memperkirakan distribusi aksi lawan $p(a_t^{-i} | h_{t-1})$ berdasarkan riwayat interaksi h_{t-1} , tanpa secara langsung dioptimalkan terhadap reward agen.

Pemisahan ini memungkinkan evaluasi model dilakukan secara independen, misalnya melalui metrik probabilistik seperti *log-likelihood* atau *cross-entropy*. Namun, dalam skenario interaksi berulang, prediksi dilakukan secara berurutan sepanjang waktu, sehingga kesalahan prediksi pada satu langkah dapat mempengaruhi langkah berikutnya, fenomena yang dikenal sebagai *compounding error*.

Selain itu, dalam pengaturan *online*, prediksi distribusi aksi lawan tidak selalu dapat diverifikasi secara langsung pada setiap langkah, terutama ketika policy agen tidak secara eksplisit mengeksplorasi prediksi tersebut. Hal ini menyebabkan adanya kesenjangan antara evaluasi berbasis akurasi prediksi dan dampaknya terhadap kinerja strategis agen.

Dengan demikian, literatur membedakan secara konseptual antara kualitas representasi belief terhadap lawan dan kualitas keputusan aksi yang dihasilkan, yang menjadi dasar bagi berbagai pendekatan dalam *opponent modelling*.

3.8.4 Risiko Propagasi Kesalahan pada Forecasting Rekursif

Pendekatan multi-step forecasting umumnya dilakukan secara rekursif, di mana prediksi pada waktu $t + 1$ digunakan sebagai input untuk memprediksi waktu $t + 2$, dan seterusnya.

Namun pendekatan ini berpotensi menimbulkan *compounding error*, di mana kesalahan kecil pada prediksi awal dapat terakumulasi dan memperbesar deviasi distribusi pada horizon yang lebih panjang (Bengio et al. 2015).

Secara formal, jika model menghasilkan distribusi prediktif $\hat{P}(a_{-i}^{t+1} | h_t)$, maka distribusi pada langkah ke- k bergantung pada distribusi hasil prediksi sebelumnya, sehingga pergeseran distribusi (*distribution shift*) dapat terjadi.

Oleh karena itu, stabilitas pelatihan menjadi aspek penting dalam pendekatan forecasting multi-langkah.

Secara umum, nilai aksi a_t didefinisikan sebagai ekspektasi akumulasi payoff terdiskonto:

$$Q(a_t) = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k u_i(a_i^{t+k}, a_{-i}^{t+k}) \right] \quad (3.31)$$

dengan $\gamma \in (0, 1)$ faktor diskonto yang ekuivalen dengan probabilitas kelanjutan permainan.

Formulasi ini menunjukkan bahwa estimasi distribusi aksi pada waktu $t + 1$ saja tidak cukup untuk mengevaluasi keputusan pada waktu t . Oleh karena itu, diperlukan prediksi lintasan masa depan melalui mekanisme multi-langkah.

3.8.5 Evaluasi Multi-Step Menggunakan Negative Log-Likelihood

Untuk mengevaluasi kualitas prediksi sekuensial lawan dalam horizon multi-step, digunakan metrik *Negative Log-Likelihood* (NLL) sebagai proper scoring rule yang konsisten terhadap estimasi distribusi probabilistik.

Misalkan pada waktu t tersedia history interaksi $h_t = \{(a_i^1, a_{-i}^1), \dots, (a_i^t, a_{-i}^t)\}$. Model pemodelan lawan menghasilkan distribusi probabilitas atas aksi lawan pada langkah berikutnya, yang dinotasikan sebagai:

$$P_{\theta}(a_{-i}^{t+1} | h_t) \quad (3.32)$$

dengan θ merepresentasikan parameter model (misalnya parameter LSTM).

Untuk prediksi multi-step sepanjang horizon H , model digunakan secara autoregresif, sehingga distribusi pada langkah ke- k bergantung pada history yang telah diperluas hingga waktu tersebut. Secara umum, probabilitas gabungan untuk urutan aksi aktual lawan $a_{-i}^{t+1:t+H}$ diberikan oleh:

$$P_\theta(a_{-i}^{t+1:t+H} \mid h_t) = \prod_{k=1}^H P_\theta(a_{-i}^{t+k} \mid \hat{h}_{t+k-1}) \quad (3.33)$$

dengan:

- a_{-i}^{t+k} : aksi aktual lawan pada waktu $t + k$,
- \hat{h}_{t+k-1} : history yang digunakan model pada langkah ke- k , yang terdiri dari history asli h_t yang diperluas secara rekursif menggunakan observasi aktual hingga waktu $t + k - 1$,
- H : panjang horizon prediksi multi-step.

Negative Log-Likelihood untuk satu segmen horizon sepanjang H kemudian didefinisikan sebagai:

$$\mathcal{L}_{\text{NLL}}^{(H)} = - \sum_{k=1}^H \log P_\theta(a_{-i}^{t+k} \mid \hat{h}_{t+k-1}) \quad (3.34)$$

Nilai NLL yang lebih kecil menunjukkan bahwa model memberikan probabilitas yang lebih tinggi terhadap aksi aktual yang benar-benar terjadi. Karena NLL merupakan proper scoring rule, metrik ini tidak hanya mengevaluasi ketepatan klasifikasi, tetapi juga kualitas kalibrasi probabilitas yang dihasilkan model.

Dalam konteks IPD dengan lawan adaptif, penggunaan NLL multi-step memungkinkan pengukuran efek *compounding error* akibat prediksi rekursif. Jika distribusi prediksi menjadi semakin bias pada horizon yang lebih panjang, maka akumulasi log-loss akan meningkat secara signifikan, mencerminkan degradasi kualitas belief seiring waktu.

BAB IV

ANALISIS DAN PERANCANGAN

4.1 Deskripsi Umum

Penelitian ini bertujuan mengembangkan algoritma pembelajaran *online* untuk skenario *Iterated Prisoner’s Dilemma* (IPD) dengan lawan adaptif dan non-stasioner. Fokus utama adalah pada pemodelan lawan yang sepenuhnya terpisah dari sinyal reward, serta eksplorasi berbasis peningkatan pemahaman terhadap dinamika lawan. Model prediksi digunakan sebagai komponen evaluasi dalam simulasi *Monte Carlo rollout* untuk mengestimasi nilai aksi agen. Pendekatan ini memisahkan sepenuhnya proses pembentukan belief dari sinyal reward, sehingga pemodelan lawan tidak ter dorong oleh bias utilitas agen. Eksplorasi berbasis epistemik dibahas sebagai kemungkinan penggunaan lanjutan dari belief state, namun tidak menjadi bagian dari implementasi utama dalam penelitian ini.

4.2 Formulasi Masalah

Penelitian ini mempertimbangkan skenario *Iterated Prisoner’s Dilemma* (IPD) dengan horizon stokastik. Pada setiap langkah waktu t , agen dan lawan masing-masing memilih aksi $a_t^{(i)} \in \{C, D\}$. Interaksi berlanjut dengan probabilitas $\gamma \in (0, 1)$ dan berhenti secara geometrik.

Lawan diasumsikan adaptif dan non-stasioner, namun algoritma internalnya tidak diketahui. Lawan dapat berupa algoritma pembelajaran daring seperti Hedge, Online Mirror Descent, atau agen berbasis reinforcement learning. Agen tidak memiliki akses terhadap parameter internal maupun fungsi utilitas lawan.

Tujuan agen adalah memaksimalkan ekspektasi reward kumulatif terhadap lawan adaptif, dengan memanfaatkan modul pemodelan lawan yang sepenuhnya terpisah dari sinyal reward.

4.3 Metodologi Penelitian

Penelitian ini bertujuan untuk membangun model *belief state* lawan dalam permainan berulang menggunakan arsitektur Long Short-Term Memory (LSTM) yang

dilatih secara *offline*, serta memanfaatkannya untuk simulasi *Monte Carlo rollout* pada tahap pengambilan keputusan.

4.3.1 Formalisasi Permainan

Permainan diformalkan sebagai permainan normal berulang dua pemain dengan: Parameter *trembling-hand* ϵ merepresentasikan probabilitas kesalahan aksi acak yang bersifat independen pada setiap tahap. Dengan probabilitas ϵ , aksi yang dipilih digantikan oleh aksi acak uniform. Parameter ini digunakan untuk memastikan bahwa proses observasi tetap eksploratif dan menghindari determinisme penuh dalam dinamika permainan.

- Himpunan aksi diskrit $A = \{1, \dots, m\}$
- Matriks payoff $P \in \mathbb{R}^{m \times m}$
- Tingkat terminasi $\gamma \in (0, 1)$
- Tingkat *trembling-hand* $\epsilon \in [0, 1]$

Pada setiap tahap t , agen dan lawan memilih aksi a_t^{agent} dan a_t^{opp} . Reward dihitung sebagai:

$$r_t^{\text{agent}}, r_t^{\text{opp}} = \phi(a_t^{\text{agent}}, a_t^{\text{opp}}, P) \quad (4.1)$$

Riwayat interaksi hingga waktu t dinotasikan sebagai:

$$H_t = \{(a_1^{\text{opp}}, a_1^{\text{agent}}, r_1^{\text{opp}}, r_1^{\text{agent}}), \dots, (a_t^{\text{opp}}, a_t^{\text{agent}}, r_t^{\text{opp}}, r_t^{\text{agent}})\} \quad (4.2)$$

Tujuan model adalah mengestimasi distribusi aksi lawan berikutnya:

$$x_{t+1} = \mathbb{P}(a_{t+1}^{\text{opp}} \mid H_t) \quad (4.3)$$

4.3.2 Arsitektur Model

Model terdiri dari dua komponen utama:

1. Encoder sekuensial berbasis LSTM
2. Jaringan feedforward untuk prediksi aksi berikutnya

Representasi Masukan Pada setiap tahap t , vektor masukan didefinisikan sebagai:

$$s_t = [a_t^{opp}, a_t^{agent}, r_t^{opp}, r_t^{agent}] \quad (4.4)$$

Seluruh parameter lingkungan seperti matriks payoff P , tingkat terminasi γ , dan tingkat *trembling-hand* ϵ digunakan dalam fungsi reward ϕ .

Dinamika Hidden State Hidden state diperbarui menggunakan LSTM:

$$h_t = \text{LSTM}(s_t, h_{t-1}) \quad (4.5)$$

dengan dimensi hidden sebesar 64 unit.

Hidden state h_t diinterpretasikan sebagai *belief state* teramortisasi yang mengaproksimasi keadaan laten strategi lawan:

$$h_t \approx q_\theta(z_t | H_t) \quad (4.6)$$

Prediksi Aksi Lawan Distribusi aksi lawan berikutnya diprediksi melalui jaringan feedforward:

$$\hat{x}_{t+1} = \text{softmax}(W h_t + b) \quad (4.7)$$

dengan $\hat{x}_{t+1} \in \Delta_m$.

Fungsi prediksi ini mengaproksimasi fungsi transisi sebenarnya:

$$x_{t+1} = f(h_t) \quad (4.8)$$

Estimasi Epistemic Uncertainty Untuk mengukur ketidakpastian epistemik model terhadap distribusi aksi lawan, digunakan pendekatan *Monte Carlo dropout* pada tahap inferensi. Dengan melakukan M kali forward pass dengan dropout aktif, diperoleh sekumpulan prediksi:

$$\{\hat{x}_{t+1}^{(m)}\}_{m=1}^M \quad (4.9)$$

Rata-rata prediksi didefinisikan sebagai:

$$\bar{x}_{t+1} = \frac{1}{M} \sum_{m=1}^M \hat{x}_{t+1}^{(m)} \quad (4.10)$$

Ketidakpastian epistemik diukur sebagai variansi prediksi:

$$U_{t+1} = \frac{1}{M} \sum_{m=1}^M \|\hat{x}_{t+1}^{(m)} - \bar{x}_{t+1}\|^2 \quad (4.11)$$

Besaran ini digunakan untuk menganalisis stabilitas belief state terhadap histori interaksi yang terbatas.

4.3.3 Fungsi Objektif

Model dilatih secara *offline* menggunakan fungsi loss *cross-entropy*:

$$\mathcal{L} = - \sum_{t=1}^T \sum_{a \in A} x_{t+1}(a) \log \hat{x}_{t+1}(a) \quad (4.12)$$

Pelatihan dilakukan dengan metode *teacher forcing*, serta *scheduled sampling* untuk mengurangi *exposure bias* pada simulasi rekursif.

4.3.4 Simulasi Monte Carlo Rollout

Setelah pelatihan selesai, model digunakan untuk simulasi multi-langkah. Distribusi prediksi digunakan untuk sampling aksi lawan:

$$a_{t+1}^{opp} \sim \hat{x}_{t+1} \quad (4.13)$$

Hidden state diperbarui secara rekursif untuk horizon k langkah. Proses ini diulang sebanyak n kali untuk memperoleh estimasi ekspektasi reward:

$$\hat{V}(H_t) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \gamma^{j-1} r_{t+j}^{(i)} \quad (4.14)$$

Estimasi nilai tersebut dihitung untuk setiap kandidat aksi $a \in A$ dengan melakukan rollout bersyarat pada aksi awal tersebut, sehingga diperoleh estimasi:

$$\hat{Q}(H_t, a) \quad (4.15)$$

Agen kemudian memilih aksi berdasarkan aturan keputusan greedy:

$$a_t^{agent} = \arg \max_{a \in A} \hat{Q}(H_t, a) \quad (4.16)$$

Diasumsikan bahwa fungsi transisi prediktif model $f : \Delta_m \rightarrow \Delta_m$ bersifat Lipschitz kontinu terhadap norma Euclidean, yaitu terdapat konstanta $L > 0$ sehingga:

$$\|f(x) - f(y)\| \leq L\|x - y\| \quad (4.17)$$

untuk setiap $x, y \in \Delta_m$. Konstanta L diestimasi secara empiris melalui maksimum rasio perubahan output terhadap perubahan input pada data validasi.

Dengan asumsi fungsi transisi Lipschitz kontinu dengan konstanta L , galat prediksi multi-langkah dibatasi oleh:

$$\|\hat{x}_{t+k} - x_{t+k}\| \leq \epsilon \sum_{i=0}^{k-1} L^i = \epsilon \frac{L^k - 1}{L - 1} \quad (4.18)$$

Sehingga horizon simulasi k dipilih agar galat tetap terkendali.

4.3.5 Pengumpulan Data

Data dikumpulkan melalui simulasi pertandingan acak antara agen dan berbagai tipe lawan baseline, meliputi:

- Strategi tetap (fixed mixed strategy)
- Tit-for-Tat
- Win-Stay Lose-Shift
- Fictitious Play
- Q-Learning

Distribusi tipe lawan dikontrol selama proses pelatihan. Dataset dibagi menjadi:

- Data pelatihan
- Data validasi
- Data pengujian (dengan variasi parameter yang tidak terlihat saat pelatihan)

Probabilitas terminasi permainan ditetapkan sebesar 0.05, sehingga panjang episode mengikuti distribusi geometrik dengan ekspektasi:

$$\mathbb{E}[T] = \frac{1}{0.05} = 20 \quad (4.19)$$

Sebanyak 10.000 episode digunakan untuk pelatihan, 2.000 untuk validasi, dan 2.000 untuk pengujian.

Distribusi tipe lawan selama pelatihan ditetapkan uniform, sehingga setiap tipe memiliki probabilitas yang sama.

4.3.6 Evaluasi

Evaluasi dilakukan berdasarkan:

1. Akurasi prediksi aksi lawan
2. Struktur representasi belief state
3. Peningkatan utilitas agen melalui Monte Carlo rollout

BAB V

HASIL DAN PEMBAHASAN

BAB VI

KESIMPULAN DAN SARAN

DAFTAR PUSTAKA

- Albrecht, S. V. dan Stone, P. (May 2018), Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems, *Artificial Intelligence* 258, pp. 66–95.
- Axelrod, R. dan Hamilton, W. D. (1981), The Evolution of Cooperation,
- Bengio, S., Vinyals, O., Jaitly, N., dan Shazeer, N. (Sept. 23, 2015), *Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks*.
- Bonanno, G., 2024, *Game theory: volume 1: basic concepts*, Kindle Direct Publishing, Place of publication not identified.
- De Weerd, H., Verbrugge, R., dan Verheij, B. (Oct. 2022), Higher-order theory of mind is especially useful in unpredictable negotiations, *Autonomous Agents and Multi-Agent Systems* 36.2, p. 30.
- Di, C., Zhou, Q., Shen, J., Wang, J., Zhou, R., dan Wang, T. (2023), The coupling effect between the environment and strategies drives the emergence of group co-operation, *Chaos, Solitons & Fractals* 176, p. 114138.
- Elhamer, Z., Suzuki, R., dan Arita, T. (2020), The effects of population size and information update rates on the emergent patterns of cooperative clusters in a large-scale social particle swarm model, *Artificial Life and Robotics* 25.1, Type: Article, pp. 149–158.
- Freire, I. T., Arsiwalla, X. D., Puigbò, J. Y., dan Verschure, P. (2023), Modeling Theory of Mind in Dyadic Games Using Adaptive Feedback Control, *Information (Switzerland)* 14.8, Type: Article.
- Gómez, A. L. d. A., Sierra, C., dan Sabater-Mir, J. (2025), Grounded predictions of teamwork as a one-shot game: A multiagent multi-armed bandits approach, *Artificial Intelligence* 341, p. 104307.
- Hernandez-Leal, P., Kartal, B., dan Taylor, M. E. (Nov. 2019), A Survey and Critique of Multiagent Deep Reinforcement Learning, *Autonomous Agents and Multi-Agent Systems* 33.6, pp. 750–797.
- Hochreiter, S. dan Schmidhuber, J. (Nov. 1, 1997), Long Short-Term Memory, *Neural Computation* 9.8, pp. 1735–1780.

- Hu, Y., Han, C., Li, H., dan Guo, T. (2023), Modeling opponent learning in multiagent repeated games, *Applied Intelligence* 53.13, Type: Article, pp. 17194–17210.
- Jin, Y., Wei, S., dan Montana, G. (Aug. 2025), Achieving collective welfare in multi-agent reinforcement learning via suggestion sharing, *Machine Learning* 114.8, p. 190.
- Li, K., Huang, W., Li, C., dan Deng, X. (2025), Exploiting a No-Regret Opponent in Repeated Zero-Sum Games; *Journal of Shanghai Jiaotong University (Science)* 30.2, Type: Article, pp. 385–398.
- Lv, M., Liu, J., Guo, B., Ding, Y., Zhang, Y., dan Yu, Z. (Sept. 2023), Inducing Coordination in Multi-Agent Repeated Game through Hierarchical Gifting Policies, *2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, 2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS), ISSN: 2155-6814, pp. 279–287.
- Nisan, N., Roughgarden, T., Tardos, É., dan Vazirani, V. V., eds. (2008), *Algorithmic game theory*, Repr., [Nachdr.], Cambridge: Cambridge Univ. Press, 754 pp.
- Perera, I., Nijs, F. de, dan Garcia, J. (2025), Learning to cooperate against ensembles of diverse opponents, *Neural Computing and Applications* 37.23, Type: Article, pp. 18835–18849.
- Qiao, X., Han, C., dan Guo, T. (Dec. 2024), O2M: Online Opponent Modeling in Online General-Sum Matrix Games, *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), pp. 358–361.
- Shoham, Y. (2009), Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations,
- Sutton, R. S. dan Barto, A. G. (2015), Reinforcement Learning: An Introduction,
- Wang, W., Wang, Y., Hao, J., dan Taylor, M. E. (2019), Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas, *ACM International Conference Proceeding Series*, Type: Conference paper.
- Zhu, L., Zhu, Y., dan Xia, C. (May 2025), Evolutionary Dynamics of Cooperation and Extortion on Networks With Fitness-Dependent Rules, *2025 Joint Internatio-*

nal Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS), 2025 Joint International Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS), ISSN: 2996-3850, pp. 1–6.

LAMPIRAN