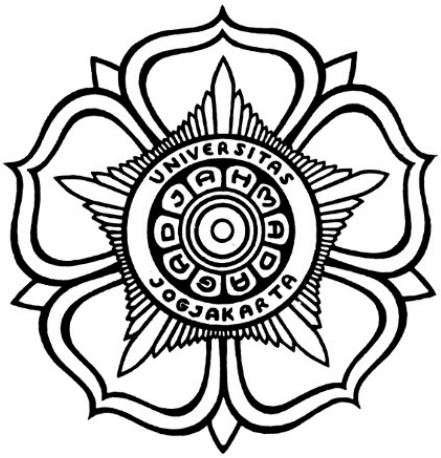
SKRIPSI

PEMODELAN LAWAN BERBASIS FORECASTING MULTI-LANGKAH YANG TERPISAH DARI OPTIMISASI REWARD PADA ITERATED PRISONER'S DILEMMA DENGAN HORIZON TERBATAS

MULTI-STEP FORECASTING-BASED OPPONENT MODELLING DECOUPLED FROM REWARD OPTIMIZATION 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 BERBASIS FORECASTING MULTI-LANGKAH YANG TERPISAH DARI OPTIMISASI REWARD PADA ITERATED PRISONER'S DILEMMA DENGAN HORIZON TERBATAS

MULTI-STEP FORECASTING-BASED OPPONENT MODELLING DECOUPLED FROM REWARD OPTIMIZATION 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 BERBASIS FORECASTING MULTI-LANGKAH YANG TERPISAH DARI OPTIMISASI REWARD 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

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

PEMODELAN LAWAN BERBASIS FORECASTING MULTI-LANGKAH YANG TERPISAH DARI OPTIMISASI REWARD 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-agen. 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 merupakan salah satu topik utama dalam multi-agent reinforcement learning (MARL) Hernandez-Leal et al. 2019. Salah satu kerangka klasik yang banyak digunakan untuk mempelajari dinamika tersebut adalah *Repeated Iterated Prisoner's Dilemma* (IPD) Axelrod dan Hamilton 1981. Dalam permainan ini, dua agen berinteraksi secara berulang tanpa komunikasi eksplisit, dan penyesuaian strategi terjadi melalui pengamatan terhadap aksi lawan pada putaran sebelumnya.

Pada IPD dengan *stochastic termination* (geometric stopping), interaksi dapat berakhir secara acak pada setiap putaran Axelrod dan Hamilton 1981. Kondisi ini menyebabkan panjang interaksi tidak pasti, meskipun memiliki nilai ekspektasi tertentu. Dalam situasi seperti ini, keputusan agen lebih bergantung pada dinamika jangka pendek dibandingkan analisis jangka panjang, karena selalu terdapat kemungkinan permainan berhenti pada langkah berikutnya Nisan et al. 2008.

Tantangan menjadi lebih besar ketika lawan bersifat adaptif atau non-stasioner, misalnya menggunakan algoritma pembelajaran daring seperti *Hedge* atau *Online Mirror Descent*, maupun pendekatan berbasis reinforcement learning Nisan et al. 2008. Dalam kondisi tersebut, agen tidak mengetahui mekanisme internal lawan dan hanya dapat mengamati urutan aksi yang terjadi selama interaksi.

Sebagian pendekatan opponent modelling dalam MARL membangun model lawan yang terintegrasi langsung dengan proses optimisasi kebijakan agen. Pembaruan model sering kali berkaitan erat dengan sinyal reward atau digunakan secara langsung untuk meningkatkan performa agen. Pendekatan ini efektif untuk tujuan performa, namun dapat menyulitkan pemisahan antara perubahan perilaku lawan dan perubahan kebijakan agen itu sendiri. Akibatnya, interpretasi terhadap dinamika strategi lawan menjadi kurang jelas.

Selain itu, banyak pendekatan berfokus pada prediksi satu-langkah (one-step prediction), yaitu memprediksi aksi lawan pada langkah berikutnya. Meskipun berguna, pendekatan ini mungkin belum cukup untuk menangkap pola strategi yang bergantung pada beberapa langkah ke depan. Dalam lingkungan dengan terminasi sto-

kastik, kemampuan melakukan prediksi multi-langkah jangka pendek (k-step forecasting) berpotensi memberikan gambaran yang lebih lengkap mengenai kecenderungan strategi lawan sebelum interaksi berakhir.

Berdasarkan kondisi tersebut, masih terdapat ruang untuk mengembangkan kerangka opponent modelling. Sebagian besar pendekatan opponent modelling dalam MARL berfokus pada inferensi tipe eksplisit, estimasi utilitas, atau pelatihan bersama berbasis reward. Dalam praktiknya, pendekatan tersebut umumnya:

1. Mengintegrasikan pembaruan model lawan secara langsung dengan objective reward agen,

2. Berorientasi pada prediksi satu langkah ke depan (one-step prediction).

Pendekatan ini efektif untuk meningkatkan performa agen, namun pemisahan antara perubahan strategi lawan dan perubahan kebijakan agen tidak selalu terlihat secara jelas. Ketika model lawan diperbarui berdasarkan sinyal reward yang sama dengan yang digunakan untuk optimisasi kebijakan, interpretasi terhadap dinamika perilaku lawan dapat menjadi kurang terpisah secara konseptual.

Selain itu, fokus pada prediksi satu langkah mungkin belum sepenuhnya menangkap pola strategi yang bergantung pada beberapa langkah interaksi. Dalam lingkungan dengan terminasi stokastik, kemampuan melakukan prediksi multi-langkah jangka pendek berpotensi memberikan informasi tambahan mengenai kecenderungan strategi lawan sebelum interaksi berakhir.

Dalam konteks ini, ketidakpastian terhadap dinamika strategi lawan dapat dipandang sebagai bagian penting dari proses pengambilan keputusan. Ketidakpastian tersebut tidak hanya berkaitan dengan variasi acak, tetapi juga dengan keterbatasan informasi yang dimiliki agen terhadap mekanisme adaptasi lawan. Oleh karena itu, selain mempertimbangkan perolehan reward, agen juga dapat mempertimbangkan bagaimana observasi baru memperbarui keyakinannya terhadap model lawan.

Berdasarkan pertimbangan tersebut, diperlukan suatu kerangka pemodelan yang (i) tidak bergantung pada asumsi eksplisit mengenai struktur algoritma internal lawan, (ii) memisahkan secara konseptual proses pembaruan belief terhadap lawan dari optimisasi reward agen, serta (iii) memungkinkan evaluasi perubahan keyakinan sebagai bagian dari analisis dinamika interaksi.

Dalam kerangka yang diusulkan, sinyal reward tetap digunakan untuk evaluasi dan seleksi kebijakan agen, namun tidak digunakan sebagai sinyal langsung dalam

pembaruan belief terhadap lawan. Dengan demikian, proses pemodelan perilaku lawan dilakukan berdasarkan observasi aksi, sementara optimisasi utilitas tetap berjalan sebagai modul terpisah.

Melalui pendekatan ini, belief terhadap dinamika lawan berfungsi tidak hanya sebagai alat prediksi, tetapi juga sebagai dasar untuk menganalisis bagaimana observasi baru memengaruhi tingkat keyakinan agen terhadap pola strategi lawan.

Dalam penelitian ini, peningkatan keyakinan (*confidence improvement*) didefinisikan sebagai perubahan terukur pada distribusi prediktif belief setelah observasi trajectory aksi agen dan lawan. Perubahan ini digunakan sebagai indikator reduksi ketidakpastian jangka pendek dalam proses interaksi.

Sebagai ilustrasi, estimasi perubahan keyakinan tersebut dapat dikaitkan dengan mekanisme seleksi aksi berbasis prinsip seperti *Upper Confidence Bound* (UCB), meskipun perancangan mekanisme seleksi aksi bukan merupakan fokus utama penelitian ini.

1.2 Rumusan Masalah

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

1. Bagaimana merancang modul pemodelan lawan yang dipisahkan dari proses optimisasi reward agen, sehingga pembaruan model dilakukan secara observasional berdasarkan trajectory aksi yang teramati?

2. Bagaimana memformulasikan dan mengimplementasikan prediksi multi-langkah jangka pendek (k-step forecasting) terhadap distribusi aksi lawan untuk menangkap dinamika strategi dalam interaksi berulang dengan terminasi stokastik?

1.3 Batasan Penelitian

Agar ruang lingkup penelitian tetap terfokus dan implementasi eksperimental terkendali, penelitian ini dibatasi pada:

1. Lingkungan *Repeated Iterated Prisoner's Dilemma* dengan *stochastic termination*.

2. Lawan adaptif yang algoritmanya tidak diketahui oleh agen, namun dibatasi pada kelas pembelajaran daring seperti *Hedge*, *Online Mirror Descent*, serta pendekatan reinforcement learning berbasis pembaruan kebijakan tanpa asumsi akses terhadap struktur internalnya.

3. Analisis difokuskan pada prediksi jangka pendek (k-step forecasting), tanpa membahas sifat asimtotik atau konvergensi jangka panjang.

4. Penelitian tidak melakukan inferensi eksplisit terhadap struktur parametrik atau bentuk algoritma internal lawan.

5. Evaluasi perubahan keyakinan dilakukan melalui perubahan distribusi prediktif belief, tanpa optimisasi eksplisit terhadap mutual information atau kriteria identifiabilitas struktural.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Merancang dan mengimplementasikan modul pemodelan belief terhadap lawan yang dipisahkan dari proses optimisasi reward agen, serta dilatih berdasarkan kesalahan prediksi (forecasting loss) terhadap aksi yang terobservasi.

2. Mengembangkan mekanisme prediksi multi-langkah jangka pendek (k-step forecasting) terhadap distribusi aksi lawan yang dikondisikan pada trajectory interaksi.

1.5 Manfaat Penelitian

Manfaat dan kontribusi yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Menyediakan formulasi opponent modelling yang dipisahkan secara struktural dari optimisasi reward agen, sehingga proses pemodelan perilaku lawan dapat dianalisis secara lebih terfokus.

2. Memperluas pendekatan prediksi dalam interaksi IPD adaptif melalui penerapan prediksi multi-langkah jangka pendek untuk menangkap dinamika strategi lawan.

3. Mengusulkan penggunaan perubahan distribusi prediktif sebagai ukuran operasional untuk mengevaluasi pembaruan keyakinan terhadap lawan dalam horizon interaksi terbatas.

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 permainaun 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.

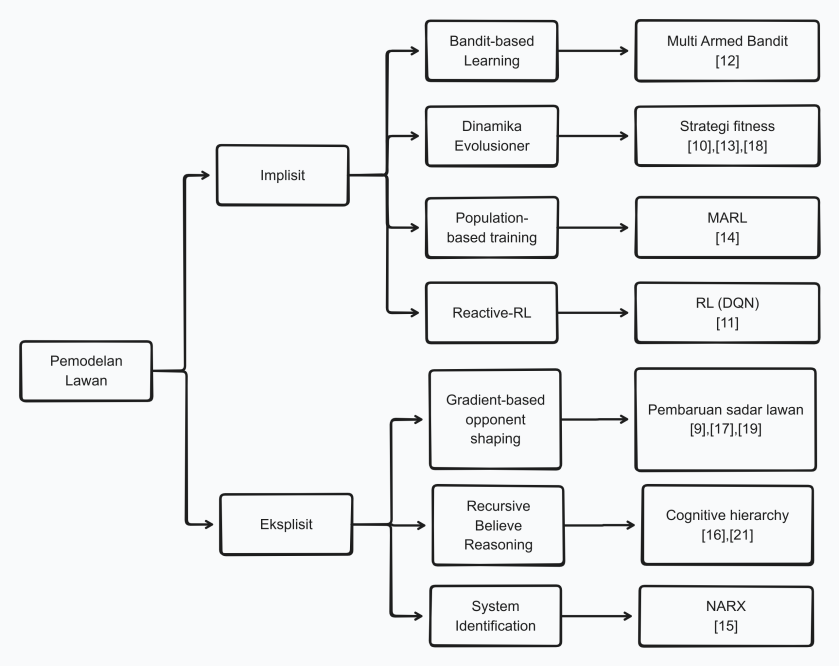
Kedua, dependensi perilaku yang dimediasi populasi terutama muncul dalam studi pada lingkungan evolusioner atau berbasis jaringan, di mana aksi individu dipengaruhi secara tidak langsung oleh dinamika agregat populasi Zhu et al. 2025; Gómez et al. 2025; Elhamer et al. 2020; Di et al. 2023; Perera et 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 autoregressive 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 sekalipun 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 eksploitasi, 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, eksploitasi 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 eksploitasi 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). Sementara 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 insentif kerja sama.
4	Teamwork Game + MA-MAB	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 suboptimal.
7	LSTM-Strategy	RNN mampu mengaproksimasi dinamika no-regret yang halus secara akurat serta memungkinkan eksploitasi yang menguntungkan pada tingkat non-stasioneritas rendah.
8	Internal Model	konvergensi lebih cepat dan stabil dibanding TD-learning standar; performa menurun 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 eksploitasi 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; meningkatkan kerja sama dalam dilema sosial.
13	Higher-order Theory of Mind (ToM k,	ToM tingkat tinggi memberikan keuntungan signifikan dalam lingkungan tidak dapat dip-

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 eksploitasi (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:

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

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:

$$G = (N, A, U) \tag{3.2}$$

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:

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

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

	C	D
C	(R, R)	(S, T)
D	(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:

$$T > R > P > S, \qquad 2R > T + S \tag{3.5}$$

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

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:

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

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:

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

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 $\gamma \in (0, 1)$:

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

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

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

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:

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

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:

$$h_t = \big((a_i^1, a_{-i}^1), (a_i^2, a_{-i}^2), \ldots, (a_i^t, a_{-i}^t)\big) \tag{3.11}$$

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:

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

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, ataupun aturan adaptasi yang digunakan oleh lawan.

Dengan demikian, meskipun seluruh pasangan aksi teramati 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:

$$b_t = P(\theta \mid h_t) \tag{3.13}$$

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:

$$h_t = \left((a_i^1, a_{-i}^1), \ldots, (a_i^t, a_{-i}^t)\right) \tag{3.14}$$

model parametrik f_θ mempelajari distribusi kondisional:

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

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:

$$h_t = \phi(W x_t + U h_{t-1} + b) \tag{3.16}$$

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:

$$b_t(a_{-i}) = P(a^t_{-i} \mid h_t) \tag{3.17}$$

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

$$\hat{b}_t = \text{softmax}(V h_t) \tag{3.18}$$

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 laten 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 laten 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

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

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

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

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

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

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:

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

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

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

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

$$h_t = o_t \odot \tanh(c_t) \tag{3.24}$$

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 laten 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 $p(a_{t+1}^{-i} \mid h_t)$. 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:

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

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:

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

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):

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

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 horizon 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 ekuilibrium 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:

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

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 bergeser 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:

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

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

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

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 $p(a_t^{-i} \mid h_{t-1})$ 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 mengeksploitasi 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 $\hat{P}(a_{-i}^{t+1} \mid h_t)$, 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:

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

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 Teacher Forcing

Pada model sekuens autoregresif seperti RNN atau LSTM, prediksi pada waktu t bergantung pada output pada waktu sebelumnya. Secara umum, dinamika sistem dapat dituliskan sebagai:

$$h_t = f(h_{t-1}, y_{t-1}; \theta) \tag{3.32}$$

$$\hat{y}_t = g(h_t; \theta) \tag{3.33}$$

dengan h_t adalah hidden state, y_{t-1} input autoregresif, $\hat{y}_t$ prediksi model, dan θ parameter yang dipelajari.

Dalam skema pelatihan standar tanpa intervensi, model menerima ground-truth y^*_{t-1} sebagai input pada setiap langkah waktu. Pendekatan ini dikenal sebagai *teacher forcing*. Dinamika selama pelatihan menjadi:

$$h_t = f(h_{t-1}, y^*_{t-1}; \theta) \tag{3.34}$$

Fungsi kerugian dihitung terhadap target aktual:

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} \ell(\hat{y}_t, y^*_t) \tag{3.35}$$

dengan $\ell(\cdot)$ merupakan fungsi loss, misalnya Mean Squared Error untuk regresi atau Cross-Entropy untuk klasifikasi.

Secara konseptual, teacher forcing mempercepat konvergensi karena distribusi input selama pelatihan identik dengan distribusi data sebenarnya. Hal ini mencegah akumulasi error pada tahap awal pembelajaran dan menghasilkan gradien yang lebih stabil.

Namun, pada fase inferensi, model tidak lagi menerima y^*_{t-1}, melainkan prediksi sebelumnya $\hat{y}_{t-1}$:

$$h_t = f(h_{t-1}, \hat{y}_{t-1}; \theta) \tag{3.36}$$

Perbedaan distribusi antara pelatihan dan inferensi ini menimbulkan *exposure bias*, di mana kesalahan kecil pada satu langkah dapat terpropagasi dan membesar pada prediksi multi-langkah.

Dengan demikian, teacher forcing efektif untuk stabilitas optimisasi, namun berpotensi menurunkan robustnes model pada skenario prediksi rekursif jangka panjang.

3.8.6 Scheduled Sampling

Meskipun *teacher forcing* mempercepat konvergensi pelatihan, pendekatan tersebut menimbulkan perbedaan distribusi antara fase pelatihan dan inferensi yang dikenal sebagai *exposure bias*. Pada saat inferensi, model tidak lagi menerima ground-truth y^*_{t-1}, melainkan prediksi sebelumnya $\hat{y}_{t-1}$, sehingga kesalahan dapat terakumulasi secara rekursif.

Untuk menjembatani kesenjangan distribusi tersebut, *scheduled sampling* mem-

perkenalkan mekanisme pencampuran antara ground-truth dan prediksi model selama pelatihan.

Pada setiap langkah waktu t, input autoregresif didefinisikan sebagai variabel acak:

$$\tilde{y}_{t-1} = \begin{cases} y^*_{t-1}, & \text{dengan probabilitas } \epsilon_k \\ \hat{y}_{t-1}, & \text{dengan probabilitas } 1 - \epsilon_k \end{cases} \tag{3.37}$$

dengan $\epsilon_k \in [0, 1]$ merupakan probabilitas penggunaan teacher forcing pada iterasi pelatihan ke-k.

Dinamika hidden state menjadi:

$$h_t = f(h_{t-1}, \tilde{y}_{t-1}; \theta) \tag{3.38}$$

dan prediksi tetap dihitung sebagai:

$$\hat{y}_t = g(h_t; \theta) \tag{3.39}$$

Nilai ϵ_k dijadwalkan menurun secara bertahap agar model secara progresif bertransisi dari rezim pelatihan berbasis ground-truth menuju rezim inferensi berbasis prediksi sendiri.

Beberapa skema penjadwalan umum adalah sebagai berikut:

Linear Decay

$$\epsilon_k = \max(\epsilon_{\min}, \epsilon_0 - \alpha k) \tag{3.40}$$

Exponential Decay

$$\epsilon_k = \epsilon_0 \gamma^k \tag{3.41}$$

Inverse Sigmoid Decay

$$\epsilon_k = \frac{\kappa}{\kappa + \exp\left(\frac{k}{\kappa}\right)} \tag{3.42}$$

dengan ϵ_0 probabilitas awal teacher forcing, α laju penurunan linear, $\gamma \in (0, 1)$ faktor peluruhan eksponensial, dan κ parameter bentuk kurva.

Dengan mekanisme ini, distribusi input selama pelatihan secara bertahap mendekati distribusi saat inferensi, sehingga model menjadi lebih stabil terhadap propagasi kesalahan pada prediksi multi-langkah.

3.8.7 Fungsi Objektif 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 $h_t = \{(a_i^1, a_{-i}^1), \ldots, (a_i^t, a_{-i}^t)\}$. Model pemodelan lawan menghasilkan distribusi probabilitas atas aksi lawan pada langkah berikutnya, yang dinotasikan sebagai:

$$P_\theta(a_{-i}^{t+1} \mid h_t) \tag{3.43}$$

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:

$$P_\theta(a_{-i}^{t+1:t+H} \mid h_t) = \prod_{k=1}^{H} P_\theta(a_{-i}^{t+k} \mid \hat{h}_{t+k-1}) \tag{3.44}$$

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:

$$\mathcal{L}_{\text{NLL}}^{(H)} = -\sum_{k=1}^{H} \log P_\theta(a_{-i}^{t+k} \mid \hat{h}_{t+k-1}) \tag{3.45}$$

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.

3.9 Optimisasi

3.9.1 Pelatihan LSTM dengan Backpropagation Through Time

Parameter LSTM $\theta = \{W, U, b\}$ dipelajari melalui minimisasi fungsi kerugian berbasis likelihood. Dalam konteks pemodelan lawan, model menghasilkan distribusi probabilitas atas aksi lawan pada setiap waktu t:

$$p_\theta(a^t_{-i} \mid h_{t-1}) \tag{3.46}$$

Diberikan urutan observasi sepanjang horizon T, fungsi kerugian total dirumuskan sebagai negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p_\theta(a^t_{-i} \mid h_{t-1}) \tag{3.47}$$

Untuk memperbarui parameter θ, gradien $\nabla_\theta \mathcal{L}$ dihitung menggunakan algoritma *Backpropagation Through Time* (BPTT).

Karena LSTM merupakan model rekuren, hidden state h_t dan cell state c_t bergantung secara rekursif pada seluruh state sebelumnya. Oleh karena itu, komputasi gradien memerlukan propagasi error dari waktu T kembali ke waktu awal melalui rantai dependensi temporal:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\partial h_t} \frac{\partial h_t}{\partial \theta} \tag{3.48}$$

Gradien pada setiap waktu tidak hanya bergantung pada kontribusi langsung terhadap loss pada waktu tersebut, tetapi juga pada pengaruh tidak langsung melalui state rekursif di masa depan. Mekanisme cell state yang bersifat aditif:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{3.49}$$

memungkinkan aliran gradien yang lebih stabil dibandingkan RNN konven-

sional, karena turunan terhadap c_{t-1} mengandung komponen linear melalui faktor f_t. Struktur ini secara signifikan mengurangi permasalahan *vanishing gradient* dalam pembelajaran dependensi jangka panjang.

Dalam praktiknya, untuk efisiensi komputasi dan stabilitas numerik, BPTT sering diterapkan dalam bentuk *truncated BPTT*, di mana propagasi gradien dibatasi pada jendela waktu sepanjang $K < T$. Parameter kemudian diperbarui menggunakan algoritma optimisasi berbasis gradien seperti Stochastic Gradient Descent (SGD) atau Adam.

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 beliefdari sinyal reward, sehingga pemodelan lawan tidak terdorong 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 membangun model prediktif aksi lawan pada permainan berulang dan memanfaatkannya untuk estimasi nilai aksi melalui simulasi multi-langkah (*Monte Carlo rollout*).

4.3.1 Setup Permainan

Permainan yang digunakan dalam penelitian ini adalah *Iterated Prisoner's Dilemma* (IPD) dua pemain dengan dua aksi diskrit:

$$A = \{C, D\} \tag{4.1}$$

di mana C menyatakan *Cooperate* dan D menyatakan *Defect*.

Matriks Payoff Struktur payoff mengikuti formulasi klasik Prisoner's Dilemma dengan parameter:

$$T > R > P > S \quad \text{dan} \quad 2R > T + S$$

Nilai numerik yang digunakan dalam penelitian ini adalah:

$$P = \begin{bmatrix} R & S \\ T & P \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix} \tag{4.2}$$

dengan:

- $T = 5$ (*Temptation*)
- $R = 3$ (*Reward*)
- $P = 1$ (*Punishment*)
- $S = 0$ (*Sucker*)

Pada setiap langkah t:

$$(a_t^{agent}, a_t^{opp}) \in \{C, D\}^2 \tag{4.3}$$

Reward agen dan lawan ditentukan langsung dari matriks payoff P.

Horizon Permainan Dua konfigurasi horizon digunakan untuk menganalisis stabilitas model:

1. **Fixed Horizon**: Panjang episode tetap $T = 100$ langkah.

2. **Geometric Termination**: Pada setiap langkah, permainan berakhir dengan probabilitas $p_{term} = 0.05$. Panjang episode mengikuti distribusi geometrik dengan ekspektasi:

$$\mathbb{E}[T] = \frac{1}{0.05} = 20 \tag{4.4}$$

Konfigurasi fixed horizon digunakan untuk analisis stabilitas prediksi multi-langkah tanpa gangguan terminasi stokastik, sedangkan konfigurasi geometrik digunakan untuk menguji robustnes model pada lingkungan dengan horizon acak.

Trembling-Hand Noise Pada eksperimen utama digunakan lingkungan deterministik ($\epsilon = 0$).

Sebagai uji robustnes, diperkenalkan *trembling-hand noise* dengan parameter $\epsilon \in \{0.01, 0.05\}$, di mana pada setiap langkah aksi yang dipilih digantikan oleh aksi acak dengan probabilitas ϵ.

Eksperimen ini bertujuan menguji stabilitas representasi *belief state* dan performa *rollout* dalam kondisi observasi yang terkontaminasi noise.

Representasi Histori Riwayat interaksi hingga waktu t dinotasikan sebagai:

$$H_t = \{(a_1^{agent}, a_1^{opp}), \ldots, (a_t^{agent}, a_t^{opp})\} \tag{4.5}$$

Tujuan model adalah mengestimasi distribusi aksi lawan berikutnya:

$$\hat{x}_{t+1} = \mathbb{P}(a_{t+1}^{opp} \mid H_t) \tag{4.6}$$

4.3.2 Arsitektur Model

Model terdiri dari dua komponen utama:

1. LSTM satu layer dengan hidden size 64 sebagai encoder temporal
2. Linear layer diikuti softmax untuk prediksi distribusi aksi lawan

Representasi Input Pada setiap langkah waktu t, model menerima vektor fitur:

$$x_t \in \mathbb{R}^d \tag{4.7}$$

yang terdiri dari:

- One-hot aksi agen a_t^{agent} (2 dimensi)
- One-hot aksi lawan a_t^{opp} (2 dimensi)
- Skor kumulatif agen hingga waktu t
- Skor kumulatif lawan hingga waktu t
- Parameter trembling-hand ϵ
- Probabilitas terminasi p_{term}

Seluruh parameter lingkungan dimasukkan pada setiap langkah waktu untuk memungkinkan generalisasi terhadap konfigurasi permainan yang berbeda.

Dinamika LSTM State internal LSTM terdiri dari hidden state h_t dan cell state c_t:

$$(h_t, c_t) = \text{LSTM}(x_t, (h_{t-1}, c_{t-1})) \tag{4.8}$$

dengan:

$$h_t \in \mathbb{R}^{64}, \quad c_t \in \mathbb{R}^{64}$$

Hidden state h_t berfungsi sebagai representasi laten histori interaksi hingga waktu t.

Prediksi Aksi Lawan Distribusi aksi lawan berikutnya diperoleh melalui:

$$\hat{x}_{t+1} = \text{softmax}(Wh_t + b) \tag{4.9}$$

dengan $\hat{x}_{t+1} \in \Delta_2$.

Prediksi Rekursif Pada tahap rollout, aksi lawan yang diprediksi $\hat{a}_{t+1}^{opp}$ dimasukkan kembali sebagai bagian dari input x_{t+1}, sehingga menghasilkan proses prediksi rekursif multi-langkah.

4.3.3 Protokol Pelatihan

Model dilatih secara *offline* menggunakan:

- Loss: cross-entropy one-step
- Optimizer: Adam
- Learning rate: 10^{-3}
- Batch size: 64
- Jumlah epoch: 50
- Early stopping berdasarkan validation loss

Pelatihan menggunakan *teacher forcing* dengan *scheduled sampling* yang meningkat secara linear dari 0 hingga 0.3 selama 30 epoch pertama.

4.3.4 Evaluasi Kandidat Aksi dan Skalabilitas Arsitektur

Pada setiap waktu t, agen membangkitkan sekumpulan kandidat aksi:

$$\mathcal{C}_t = \{c^{(1)}, c^{(2)}, \ldots, c^{(N)}\} \tag{4.10}$$

dengan N bersifat arbitrer. Arsitektur yang digunakan tidak membatasi jumlah kandidat secara struktural, sehingga evaluasi aksi berskala linear terhadap N.

Untuk setiap kandidat $c^{(i)}$, dilakukan estimasi nilai menggunakan simulasi Monte Carlo rollout sepanjang horizon k dengan n simulasi independen:

$$\hat{Q}(H_t, c^{(i)}) = \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=1}^{k} \gamma^{\tau-1} r_{t+\tau}^{(j,i)} \tag{4.11}$$

di mana H_t adalah riwayat hingga waktu t, dan $r_{t+\tau}^{(j,i)}$ adalah reward pada langkah τ dari rollout ke-j untuk kandidat $c^{(i)}$.

Estimasi Ketidakpastian dengan Monte Carlo Dropout Untuk menangkap ketidakpastian epistemik, digunakan pendekatan *Monte Carlo dropout*. Dropout tetap aktif saat inferensi dan dilakukan sebanyak M forward pass independen.

Untuk setiap kandidat $c^{(i)}$, diperoleh:

$$\hat{Q}^{(m)}(H_t, c^{(i)}), \quad m = 1, \ldots, M \tag{4.12}$$

Nilai ekspektasi diperkirakan sebagai:

$$\bar{Q}(c^{(i)}) = \frac{1}{M} \sum_{m=1}^{M} \hat{Q}^{(m)}(H_t, c^{(i)}) \tag{4.13}$$

Sedangkan variansi antar forward pass:

$$\mathrm{Var}_Q(c^{(i)}) = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{Q}^{(m)}(H_t, c^{(i)}) - \bar{Q}(c^{(i)}) \right)^2 \tag{4.14}$$

digunakan sebagai indikator ketidakpastian epistemik terhadap estimasi nilai aksi.

Pemilihan Aksi Aksi dipilih secara greedy berdasarkan estimasi nilai rata-rata:

$$a_t^{agent} = \arg\max_{c^{(i)} \in \mathcal{C}_t} \bar{Q}(c^{(i)}) \tag{4.15}$$

Sebagai alternatif, dapat digunakan kriteria *risk-sensitive*:

$$a_t^{agent} = \arg\max_{c^{(i)} \in \mathcal{C}_t} \left(\bar{Q}(c^{(i)}) - \lambda \sqrt{\mathrm{Var}_Q(c^{(i)})} \right) \tag{4.16}$$

dengan $\lambda \geq 0$ sebagai parameter aversi risiko.

Parameter Eksperimen Secara teoritis, jumlah kandidat N bersifat arbitrer dan tidak membatasi struktur model. Namun dalam eksperimen ini digunakan:

- Jumlah kandidat aksi: $N = 20$
- Jumlah rollout per kandidat: $n = 50$
- Horizon simulasi: $k = 10$
- Jumlah forward pass MC dropout: $M = 20$

Pemilihan $N = 20$ didasarkan pada kompromi antara stabilitas estimasi Monte Carlo (dengan error $\mathcal{O}(1/\sqrt{N})$) dan efisiensi komputasi, mengingat kompleksitas total evaluasi berskala:

$$\mathcal{O}(N \cdot n \cdot k \cdot M) \tag{4.17}$$

Dengan demikian, nilai $N = 20$ cukup untuk menghasilkan estimasi yang stabil tanpa meningkatkan beban komputasi secara berlebihan, sementara arsitektur tetap mendukung jumlah kandidat yang lebih besar pada skenario lanjutan.

4.3.5 Pengumpulan Data

Dataset dikumpulkan melalui simulasi pertandingan melawan berbagai tipe lawan:

- Fixed mixed strategy
- Tit-for-Tat
- Win-Stay Lose-Shift
- Fictitious Play
- Q-Learning

Distribusi tipe lawan selama pelatihan ditetapkan uniform.
Jumlah episode:

- 10.000 pelatihan
- 2.000 validasi
- 2.000 pengujian

Set pengujian mencakup variasi parameter yang tidak terlihat saat pelatihan untuk menguji generalisasi.

4.3.6 Evaluasi

Evaluasi dilakukan secara bertingkat untuk memisahkan kontribusi dari model prediktif, mekanisme rollout, estimasi ketidakpastian, serta kualitas pengambilan keputusan dalam interaksi tertutup. Pendekatan ini memastikan bahwa setiap komponen sistem dianalisis secara terpisah sebelum dievaluasi secara end-to-end.

Lingkungan Evaluasi dan Tipe Lawan Evaluasi dilakukan dalam lingkungan *Iterated Prisoner's Dilemma* (IPD) dengan berbagai tipe lawan untuk menguji robustnes model terhadap dinamika yang berbeda. Setiap tipe lawan merepresentasikan kelas strategi dengan karakteristik stasioner maupun adaptif.

- **Fixed Mixed Strategy**
  Lawan memilih aksi kooperasi dengan probabilitas tetap $p \in [0, 1]$ dan defeksi dengan probabilitas $1 - p$. Strategi ini bersifat stasioner dan tidak bergantung pada riwayat permainan.

- **Tit-for-Tat (TFT)**
  Lawan memulai dengan kooperasi, kemudian meniru aksi terakhir agen:

  $$a_t^{opp} = a_{t-1}^{agent}.$$

  Strategi ini deterministik dan reaktif.

- **Win-Stay Lose-Shift (WSLS)**
  Lawan mempertahankan aksi sebelumnya jika reward tinggi, dan mengganti aksi jika reward rendah. Strategi ini bergantung pada payoff sebelumnya.

- **Fictitious Play**
  Lawan mengestimasi distribusi aksi agen berdasarkan frekuensi historis dan memilih aksi terbaik terhadap distribusi tersebut. Strategi ini adaptif namun berbasis estimasi frekuensi.

- **Q-Learning Agent**
  Lawan merupakan agen pembelajaran berbasis Q-learning yang memperbarui nilai aksi secara iteratif selama episode. Strategi ini bersifat non-stasioner dan adaptif terhadap dinamika permainan.

Setiap tipe lawan dievaluasi dalam sejumlah episode independen. Parameter strategi (misalnya probabilitas pada fixed mixed strategy atau learning rate pada Q-learning) ditetapkan konstan selama satu episode namun dapat divariasikan antar eksperimen untuk menguji robustnes agen terhadap perubahan dinamika lingkungan.

Evaluasi Model Prediktif (Offline) Evaluasi pertama dilakukan pada tahap offline untuk mengukur kualitas model LSTM sebagai prediktor aksi lawan secara lokal (one-step).

Metrik yang digunakan:

- **One-step accuracy**
- **Cross-entropy loss**
- **Entropy prediktif** sebagai indikator tingkat keyakinan model
- Analisis perbandingan dengan dan tanpa *scheduled sampling*

Tahap ini bertujuan memastikan bahwa model memiliki kapasitas representasi yang memadai sebelum digunakan dalam rollout multi-langkah.

Evaluasi Stabilitas Multi-Langkah (Open-loop) Pada tahap ini model dievaluasi tanpa intervensi agen, dengan melakukan prediksi rekursif hingga horizon k.

Dievaluasi:

- Akurasi prediksi sebagai fungsi horizon k
- Pertumbuhan error terhadap peningkatan horizon
- Divergensi distribusi prediktif terhadap distribusi ground-truth
- Pertumbuhan variansi prediktif:

$$\mathrm{Var}(x_{t+k}) \text{ sebagai fungsi } k$$

Analisis ini mengukur stabilitas representasi laten serta tingkat akumulasi error akibat self-conditioning.

Evaluasi Estimasi Ketidakpastian Ketidakpastian epistemik dievaluasi menggunakan Monte Carlo dropout dengan M forward pass independen.

Untuk setiap prediksi diperoleh:

- Rata-rata probabilitas prediksi
- Variansi antar forward pass
- Entropy distribusi prediktif

Selain itu dianalisis:

- Korelasi antara variansi prediktif dan error aktual
- Pertumbuhan ketidakpastian terhadap peningkatan horizon

Tujuannya adalah memverifikasi bahwa estimasi ketidakpastian bersifat informatif dan berkorelasi dengan kesalahan prediksi.

Evaluasi Kualitas Perencanaan (Planning Quality) Evaluasi ini menilai stabilitas estimasi nilai aksi $\hat{Q}$.

Dianalisis sensitivitas terhadap parameter simulasi:

- Horizon rollout k
- Jumlah rollout per kandidat n
- Jumlah kandidat aksi N
- Jumlah forward pass MC dropout M

Diperiksa konvergensi estimator Monte Carlo dengan memperhatikan:

$$\mathrm{Var}(\hat{Q}) \propto \frac{1}{n}$$

serta stabilitas estimasi terhadap variasi M.

Tahap ini memastikan bahwa performa agen bukan artefak dari konfigurasi simulasi tertentu.

Evaluasi Performa Agen (Closed-loop) Evaluasi akhir dilakukan dalam interaksi tertutup melalui simulasi episode permainan penuh.

Metrik yang digunakan:

- Rata-rata reward per episode
- Distribusi reward
- Win-rate terhadap baseline
- Tingkat kooperasi sepanjang episode
- Stabilitas pola interaksi antar waktu

Baseline Performa agen dibandingkan dengan beberapa baseline:

- Agen greedy one-step (tanpa rollout)
- Agen rollout tanpa estimasi ketidakpastian
- Agen heuristic reaktif
- Agen acak (random policy)

Perbandingan ini bertujuan mengisolasi kontribusi rollout multi-langkah serta estimasi ketidakpastian terhadap peningkatan performa agen.

Ablation Study Dilakukan studi ablasi untuk mengevaluasi pengaruh setiap komponen sistem:

- Tanpa scheduled sampling
- Tanpa Monte Carlo dropout
- Horizon tetap ($k = 1$)
- Variasi jumlah kandidat aksi N

Ablasi ini membantu mengidentifikasi komponen yang paling berkontribusi terhadap stabilitas dan performa akhir agen.

BAB V
JADWAL PENELITIAN

5.1 Jadwal Penelitian

Penelitian ini direncanakan berlangsung selama enam bulan dengan pembagian tahapan yang sistematis sesuai dengan alur metodologi yang telah dijelaskan pada Bab 4. Penyusunan jadwal dilakukan berdasarkan dependensi antar komponen sistem, dimulai dari pembangunan lingkungan simulasi, implementasi model prediktif, hingga evaluasi tertutup dan studi ablasi.

5.1.1 Tahapan Penelitian

Secara umum, tahapan penelitian dibagi menjadi sebelas fase utama sebagai berikut:

1. **Perancangan dan Validasi Lingkungan IPD**

   Tahap awal mencakup implementasi lingkungan *Iterated Prisoner's Dilemma* (IPD), termasuk:

   - Implementasi matriks payoff
   - Implementasi fixed horizon ($T = 100$)
   - Implementasi geometric termination ($p_{term} = 0.05$)
   - Implementasi trembling-hand noise ($\epsilon \in \{0, 0.01, 0.05\}$)
   - Validasi distribusi panjang episode
   - Pengujian konsistensi reward dan terminasi

   Tahap ini memastikan bahwa seluruh eksperimen dilakukan pada lingkungan yang tervalidasi dan bebas dari kesalahan logika.

2. **Implementasi Tipe Lawan**

   Meliputi pembangunan berbagai tipe lawan yang digunakan dalam pengumpulan data dan evaluasi, yaitu:

   - Fixed mixed strategy

- Tit-for-Tat
- Win-Stay Lose-Shift
- Fictitious Play
- Q-Learning agent

Dilakukan pula validasi bahwa agen adaptif benar-benar menunjukkan dinamika non-stasioner selama episode.

3. **Pengumpulan dan Validasi Dataset**

   Tahap ini mencakup:

   - Simulasi 10.000 episode pelatihan
   - Simulasi 2.000 episode validasi
   - Simulasi 2.000 episode pengujian
   - Verifikasi distribusi tipe lawan uniform
   - Pemeriksaan data leakage

4. **Implementasi Model LSTM**

   Meliputi:

   - Implementasi arsitektur LSTM satu layer (hidden size 64)
   - Integrasi dropout untuk Monte Carlo dropout
   - Implementasi scheduled sampling ($0 \rightarrow 0.3$)
   - Implementasi training pipeline (Adam, early stopping)

5. **Pelatihan Model Prediktif**

   Meliputi:

   - Pelatihan model dengan teacher forcing
   - Monitoring training dan validation loss
   - Penyimpanan model terbaik berdasarkan validation loss

6. **Evaluasi Model Prediktif (Offline)**

   Evaluasi one-step prediction dengan metrik:

- Accuracy
- Cross-entropy loss
- Entropy prediktif
- Analisis pengaruh scheduled sampling

7. **Evaluasi Stabilitas Multi-Langkah (Open-loop)**

   Meliputi:

   - Prediksi rekursif hingga horizon $k = 10$
   - Analisis pertumbuhan error terhadap horizon
   - Analisis KL-divergence
   - Analisis pertumbuhan variansi prediktif

8. **Implementasi Monte Carlo Rollout Planning**

   Meliputi:

   - Implementasi estimator $\hat{Q}$
   - Integrasi Monte Carlo dropout ($M = 20$)
   - Validasi kompleksitas $\mathcal{O}(N \cdot n \cdot k \cdot M)$
   - Analisis konvergensi estimator terhadap n dan M

9. **Evaluasi Closed-loop Agen**

   Meliputi:

   - Simulasi episode penuh
   - Evaluasi reward rata-rata
   - Evaluasi win-rate
   - Evaluasi tingkat kooperasi
   - Uji robustnes terhadap noise dan horizon geometrik

10. **Studi Ablasi**

    Meliputi:

- Tanpa scheduled sampling
- Tanpa Monte Carlo dropout
- Horizon $k = 1$
- Variasi jumlah kandidat aksi N

11. **Analisis Statistik dan Penulisan Laporan**

    Tahap akhir mencakup:

    - Uji signifikansi statistik
    - Analisis interval kepercayaan
    - Visualisasi hasil eksperimen
    - Penyusunan Bab 4 dan Bab 5 final
    - Revisi berdasarkan masukan pembimbing

5.1.2 Rencana Waktu Pelaksanaan

Rencana waktu pelaksanaan penelitian selama enam bulan ditunjukkan pada Tabel berikut.

Tabel 5.1 Rencana Jadwal Penelitian (2 Bulan / 8 Minggu)

Kegiatan	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Perancangan lingkungan IPD	✓	✓						
Implementasi tipe lawan	✓	✓						
Pengumpulan dan validasi dataset		✓	✓					
Implementasi model LSTM			✓	✓				
Pelatihan model				✓	✓			
Evaluasi model offline (one-step)					✓			
Evaluasi open-loop multi-step					✓	✓		
Implementasi Monte Carlo rollout						✓	✓	
Evaluasi closed-loop agen							✓	
Studi ablasi							✓	
Analisis statistik dan visualisasi								✓
Penulisan dan revisi laporan	✓	✓	✓	✓	✓	✓	✓	✓

5.1.3 Dependensi dan Risiko

Beberapa dependensi penting dalam penelitian ini:

- Implementasi rollout planning bergantung pada stabilitas model prediktif pada evaluasi open-loop.
- Evaluasi closed-loop tidak dapat dilakukan sebelum estimator nilai $\hat{Q}$ tervalidasi.
- Studi ablasi dilakukan setelah konfigurasi utama sistem stabil.

Risiko utama penelitian meliputi:

- Akumulasi error prediksi pada horizon panjang
- Kompleksitas komputasi akibat $\mathcal{O}(N \cdot n \cdot k \cdot M)$
- Ketidakstabilan model pada lawan non-stasioner

Strategi mitigasi meliputi pengujian bertahap (offline $\rightarrow$ open-loop $\rightarrow$ closed-loop), monitoring konvergensi estimator, serta pembatasan parameter eksperimen untuk menjaga efisiensi komputasi.

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 cooperation, *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