IMPLEMENTASI *AGENT DEEP REINFORCEMENT LEARNING* MENGGUNAKAN *PROXIMAL POLICY OPTIMIZATION* DENGAN *REWARD SHAPING* DAN EVALUASI MULTI-METRIK PADA *ENVIRONMENT HUMANOID*

TUGAS AKHIR

Disusun Untuk Memenuhi Syarat Kelulusan Program Strata I
pada Sekolah Tinggi Manajemen Informatika dan Komputer
(STMIK) Palangkaraya

OLEH

RIO JULIANNUR SAPUTRA
C2255201075
PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
2026

IMPLEMENTASI *AGENT DEEP REINFORCEMENT LEARNING* MENGGUNAKAN *PROXIMAL POLICY OPTIMIZATION* DENGAN *REWARD SHAPING* DAN EVALUASI MULTI-METRIK PADA *ENVIRONMENT HUMANOID*

TUGAS AKHIR

Disusun Untuk Memenuhi Syarat Kelulusan Program Strata I
pada Sekolah Tinggi Manajemen Informatika dan Komputer
(STMIK) Palangkaraya

OLEH

RIO JULIANNUR SAPUTRA
C2255201075
PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
2026

PERSETUJUAN

IMPLEMENTASI *AGENT DEEP REINFORCEMENT LEARNING* MENGGUNAKAN *PROXIMAL POLICY OPTIMIZATION* DENGAN *REWARD SHAPING* DAN EVALUASI MULTI-METRIK PADA *ENVIRONMENT HUMANOID*

Tugas Akhir Ini Telah Disetujui dan Disahkan

Pembimbing I,	Pembimbing II,
Lili Rusdiana, M.Kom.	Catharina Elmayantie, M.Pd.
NIK. 198707282011007	NIK. 197610252015003

Mengetahui	Mengesahkan
Ketua STMIK Palangkaraya,	Ketua Prodi Teknik Informatika,
Ir. Hj. Siti Maryamah, M.M	Ir. Abdul Hadi, S.T., M.Kom
NIK. 196902021995004	NIK. 198505102021101

PENGESAHAN

IMPLEMENTASI *AGENT DEEP REINFORCEMENT LEARNING* MENGGUNAKAN *PROXIMAL POLICY OPTIMIZATION* DENGAN *REWARD SHAPING* DAN EVALUASI MULTI-METRIK PADA *ENVIRONMENT HUMANOID*

Tugas Akhir ini telah Diujikan, Dinilai, dan Disahkan
Oleh Tim Penguji pada Tanggal 02 Mei 2026

Tim Penguji Tugas Akhir :

1. Veny Cahya Hardita, M.Kom ..... 
   Ketua

2. Rudini, M.Pd ..... 
   Sekretaris

3. Dr. Maura Widyaningsih, S.Kom., M.Cs .....
   Anggota

4. Lili Rusdiana, M.Kom ..... 
   Anggota

5. Catharina Elmayantie, M.Pd ..... 
   Anggota

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama Mahasiswa : Rio Juliannur Saputra

N I M : C2255201075

menyatakan bahwa Tugas Akhir dengan judul :

IMPLEMENTASI *AGENT DEEP REINFORCEMENT LEARNING* MENGGUNAKAN *PROXIMAL POLICY OPTIMIZATION* DENGAN *REWARD SHAPING* DAN EVALUASI MULTI-METRIK PADA *ENVIRONMENT HUMANOID*

adalah hasil karya saya dan bukan merupakan duplikasi sebagian atau seluruhnya dari karya orang lain, kecuali bagian yang sumber informasi dicantumkan.

Pernyataan ini dibuat dengan sebenar-benarnya secara sadar dan bertanggungjawab dan saya bersedia menerima sanksi pembatalan Tugas Akhir apabila terbukti melakukan duplikasi terhadap Tugas Akhir atau karya ilmiah lain yang sudah ada.

Palangka Raya, 02 Mei 2026

Yang Membuat Pernyataan,



Rio Juliannur Saputra

MOTTO DAN PERSEMBAHAN

"Machines can learn, but it is the human who must first dare to teach."
— Andrew Ng, AI Researcher & Co-founder of Coursera

"Setiap kegagalan adalah sinyal agar kita menyesuaikan langkah, bukan berhenti melangkah." — Orisinal

Dengan segenap rasa syukur dan kerendahan hati, karya sederhana ini kupersembahkan kepada:

- **Allah Subhanahu wa Ta'ala,** atas segala rahmat, kemudahan, dan kekuatan yang selalu diberikan di setiap langkah perjalanan ini.
- **Kedua orang tuaku tercinta,** yang tidak pernah lelah mendoakan, tidak pernah berhenti berkorban, dan selalu menjadi alasan terkuat untuk terus bangkit ketika semangat mulai goyah. Setiap baris dalam karya ini adalah bukti cinta yang ingin kubalas, meski tidak akan pernah cukup.
- **Kedua kakakku tersayang,** yang selalu hadir dengan perlindungan, perhatian, dan dorongan yang tulus. Terima kasih telah menjadi teladan sekaligus penyemangat, dan tidak pernah berhenti percaya bahwa adikmu ini mampu menyelesaikan apa yang telah dimulai.
- **Seluruh keluarga besar,** atas doa-doa yang terucap diam-diam dan semangat yang tak pernah padam.
- **Sahabat dan teman-teman seperjuangan di STMIK Palangkaraya,** terima kasih untuk tawa, Ilmu, dan perjuangan yang kita lalui bersama.

INTISARI

Rio Juliannur Saputra, C2255201075, 2026. *Implementasi agent Deep Reinforcement Learning menggunakan Proximal Policy Optimization dengan reward Shaping dan evaluasi Multi-metrik pada Environment Humanoid,* Pembimbing I Lili Rusdiana, M.Kom., Pembimbing II Catharina Elmayantie, M.Pd.

Penelitian ini mengembangkan agen *Deep Reinforcement Learning* berbasis *Proximal Policy Optimization* (PPO) yang ditingkatkan dengan mekanisme *Reward Shaping* berbasis domain untuk meningkatkan stabilitas dan efisiensi pelatihan pada lingkungan simulasi Humanoid. PPO standar tanpa modifikasi (baseline) terbukti suboptimal dalam eksperimen ini, hanya mencapai *mean reward* sebesar 1.215,92 dengan rata-rata panjang episode sebesar 222,8 langkah, menunjukkan tren ketidakstabilan yang konsisten, serta gagal menghasilkan nilai stabilitas torso yang terukur. Lebih lanjut, evaluasi kinerja agen menggunakan satu metrik tunggal dinilai tidak memadai untuk mendeskripsikan kualitas agen secara komprehensif. Oleh karena itu, penelitian ini membangun agen PPO yang ditingkatkan menggunakan *Reward Shaping* dan mengevaluasinya melalui pendekatan multi-metrik pada lingkungan *Humanoid-v5*.

Penelitian ini menggunakan pendekatan eksperimen dan pengembangan (R&D) dengan dua skenario pelatihan: PPO Standar sebagai pembanding dasar (*baseline*), dan PPO yang ditambahkan *Reward Shaping*. Kedua skenario dilatih selama 50 juta langkah pada simulasi robot Humanoid menggunakan platform MuJoCo. *Reward Shaping* dirancang dengan tiga komponen tambahan yaitu stabilitas postur tubuh, efisiensi penggunaan energi, dan ketahanan gerak. Kinerja agen dievaluasi menggunakan lima ukuran utama: *reward* rata-rata, tingkat keberhasilan, panjang episode, stabilitas torso, dan efisiensi energi.

Hasil penelitian menunjukkan bahwa *PPO* dengan *Reward Shaping* secara konsisten mengungguli *PPO* Standar pada seluruh metrik. *Mean reward* meningkat sebesar 611,0% dari 1.215,92 menjadi 8.641,51, episode *length* meningkat 320,2% dari 222,8 menjadi 934,96 langkah, torso stability hanya terukur pada skenario *Reward Shaping* dengan nilai 1,363, dan efisiensi energi meningkat 46,0%. *PPO* dengan *Reward Shaping* juga mencapai konvergensi lebih cepat pada *timestep* ke-38.889.108 dibandingkan *PPO* Standar pada *timestep* ke-44.200.000.

Kata kunci : *Deep Reinforcement Learning, Environment Humanoid Evaluasi Multi-Metrik, Humanoid, Proximal Policy Optimization*

ABSTRACT

Rio Juliannur Saputra, C2255201075, 2026. *Implementation of a Deep Reinforcement Learning Agent Using Proximal Policy Optimization with Reward Shaping and Multi-Metric Evaluation on the Humanoid Environment, Advisor* I Lili Rusdiana, M.Kom., *Advisor* II Catharina Elmayantie, M.Pd.

This study develops a Proximal Policy Optimization (PPO)-based Deep Reinforcement Learning agent enhanced with a domain-based Reward Shaping mechanism to improve training stability and efficiency on the Humanoid simulation environment. Standard PPO without modification (baseline) is proven suboptimal, in this experiment, it only achieved a mean reward of 1,215.92 with an average episode length of 222.8 steps, exhibiting a consistent instability trend and failing to produce a measurable torso stability value. Furthermore, evaluating agent performance using only a single metric is insufficient to comprehensively describe agent quality. This study therefore builds an enhanced PPO agent using Reward Shaping and evaluates it through a multi-metric approach on the Humanoid-v5 Environment.

The method employed is computational experimentation with a Research and Development (R&D) approach. Two training scenarios were conducted separately, namely Standard PPO as the baseline and PPO with Reward Shaping, both trained for 50,000,000 timesteps using the Stable-Baselines3 framework on the MuJoCo simulator through the Gymnasium interface. Reward Shaping was designed with three domain-based components, namely postural stability, energy efficiency, and movement endurance. Data collection was carried out through a literature study to establish the theoretical foundation, as well as automatic trajectory data logging during the training process using the logging feature provided by the Stable-Baselines3 framework. Evaluation was performed using five main metrics, namely mean reward, success rate, episode length, torso stability, and energy efficiency.

The results show that PPO with Reward Shaping consistently outperformed Standard PPO across all metrics. Mean reward increased by 611.0% from 1,215.92 to 8,641.51, episode length increased by 320.2% from 222.8 to 934.96 steps, torso stability was only measurable in the Reward Shaping scenario with a value of 1.363, and energy efficiency improved by 46.0%. PPO with Reward Shaping also achieved faster convergence at timestep 38,889,108 compared to Standard PPO at timestep 44,200,000.

***Keywords**: Deep Reinforcement Learning, Energy Efficiency, Humanoid Environment, Multi-Metric Evaluation, Proximal Policy Optimization*

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai sivitas akademik Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya, saya yang bertanda tangan di bawah ini:

Nama : Rio Juliannur Saputra

NIM : C2255201075

Program Studi : Teknik Informatika

Jenis Karya : Laporan Tugas Akhir

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya, Hak Bebas Royalti Non Eksklusif (*Non-exclusive Royalty-Free Right*) atas karya ilmiah saya yang berjudul:

IMPLEMENTASI *AGENT DEEP REINFORCEMENT LEARNING* MENGGUNAKAN *PROXIMAL POLICY OPTIMIZATION* DENGAN *REWARD SHAPING* DAN EVALUASI MULTI-METRIK PADA *ENVIRONMENT HUMANOID*

Dengan Hak Bebas Royalti Non Eksklusif ini Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Palangka Raya
Pada tanggal : 02 Mei 2026

Yang menyatakan



Rio Juliannur Saputra

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah Subhanahu wa Ta'ala atas segala rahmat, taufik, dan hidayah-Nya sehingga penulis dapat menyelesaikan tugas akhir yang berjudul "Implementasi Agent Deep Reinforcement Learning menggunakan Proximal Policy Optimization dengan Reward Shaping dan Evaluasi Multi-Metrik pada Environment Humanoid" ini dengan baik. Tugas akhir ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata I pada Program Studi Teknik Informatika di Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya.

Dalam proses penyusunan tugas akhir ini, penulis memperoleh berbagai bantuan, dukungan, serta bimbingan dari banyak pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, dengan penuh rasa hormat penulis menyampaikan terima kasih kepada:

1. Lili Rusdiana, M.Kom. selaku Dosen Pembimbing I, yang telah meluangkan waktu dan tenaga untuk memberikan bimbingan teknis, arahan, serta motivasi selama proses penyusunan tugas akhir ini sehingga penelitian dapat berjalan dengan terarah dan terselesaikan dengan baik.

2. Catharina Elmayantie, M.Pd. selaku Dosen Pembimbing II, yang telah memberikan bimbingan dalam hal penulisan ilmiah, saran, serta masukan yang sangat membantu penulis dalam menyusun laporan tugas akhir ini hingga layak untuk dipertanggungjawabkan secara akademis.

Kritik dan saran yang bersifat membangun sangat penulis harapkan demi perbaikan ke depannya. Semoga karya ini dapat memberikan manfaat bagi pembaca, menjadi referensi yang berguna bagi penelitian selanjutnya, serta berkontribusi pada perkembangan ilmu pengetahuan, khususnya di bidang kecerdasan buatan dan deep reinforcement learning.

Palangka Raya, 02 Mei 2026

Penulis

DAFTAR ISI

PENGESAHAN ..... **iv**
PERNYATAAN KEASLIAN ..... **v**
MOTTO DAN PERSEMBAHAN ..... **vi**
INTISARI ..... **vi**
ABSTRACT ..... **viii**
KATA PENGANTAR ..... **x**
DAFTAR ISI **11**
DAFTAR TABEL **13**
DAFTAR GAMBAR **14**
DAFTAR LAMPIRAN **15**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang Masalah 1
1.2 Perumusan Masalah 4
1.3 Batasan Masalah 4
1.4 Tujuan dan Manfaat 4
1.5 Sistematika Penulisan 5
BAB II TINJAUAN PUSTAKA **7**
2.1 Kajian Teori 7
2.1.1 Konsep Dasar *Reinforcement Learning* (RL) 7
2.1.2 Markov Decision Process (MDP) 8
2.1.3 *Deep Reinforcement Learning* (*DRL*) 9
2.1.4 *Proximal Policy Optimization* (*PPO*) 9
2.1.4 *Reward Shaping* 10
2.1.5 Evaluasi Multi-Metrik 11
2.1.6 *Environment Humanoid* 12
2.1.7 Stable-*Baseline* s3 (SB3) 13
2.1.7 Metode *Research and Development* (R&D) 13
2.2 Penelitian yang Relevan 16
BAB III METODOLOGI PENELITIAN **19**
3.1 Tinjauan Umum 19

3.2 Jenis Penelitian ……20
3.3 Desain Penelitian …… 24
3.4 Instrumen Penelitian ……26
3.5 Analisis Kebutuhan ……27
3.5.1 Kebutuhan Perangkat Keras ……27
3.5.2 Kebutuhan Perangkat Lunak ……27
3.5.1 Analisis Data ……27
3.5.2 Analisis Proses ……28
BAB IV HASIL DAN PEMBAHASAN 33
4.1 Hasil …… 33
4.1.1 Implementasi ……33
4.1.2 Pengujian ……..37
4.2 Pembahasan ……42
BAB V KESIMPULAN DAN SARAN 46
5.1 Kesimpulan …… 46
5.2 Saran ……47
DAFTAR PUSTAKA 49

DAFTAR TABEL

Tabel 2. 1 Penelitian Yang Relevan 16
Tabel 3. 1 Kebutuhan Perangkat Keras 27
Tabel 3. 2 Kebutuhan Perangkat Lunak 27
Tabel 4. 1 Konfigurasi *hyperparameter PPO* 33
Tabel 4. 2 Komponen dan Koefisien *Reward Shaping* 35
Tabel 4. 3 Perbandingan Lima Metrik Evaluasi Utama 40
Tabel 4. 4 Perkembangan Performa *Agent* Selama Proses Pelatihan 41

DAFTAR GAMBAR

Gambar 3. 1 Flowchart Proses Pelatihan *Agent PPO* dengan *Reward Shaping* ... 29
Gambar 4. 1Hasil Pelatihan *Agent PPO* + *Reward Shaping* pada *Environment Humanoid*-v5 36
Gambar 4. 2 Perbandingan Kurva *Mean Reward PPO Baseline* vs *PPO* + *Reward Shaping*..... 37
Gambar 4. 3 Perbandingan Kurva Episode *Length PPO Baseline* vs *PPO* + *Reward Shaping* 38
Gambar 4. 4 Perbandingan Nilai Akhir *Reward* dan Episode *Length* (Diagram Batang) 39
Gambar 4. 5 Perbandingan Kecepatan Konvergensi *PPO Baseline* vs *PPO* + *Reward Shaping* 39

DAFTAR LAMPIRAN

Lampiran 1. Surat tugas pembimbing tugas akhir
Lampiran 2. Lembar konsultasi bimbingan Tugas Akhir
Lampiran 3. Surat tugas penguji seminar
Lampiran 4. Poster Penelitian
Lampiran 5. *Listing* Program

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Kecerdasan buatan (Artificial Intelligence/AI) telah menjadi salah satu bidang yang berkembang pesat, terutama melalui pendekatan Machine *Learning* (ML). Salah satu cabang yang semakin mendapat adalah *Reinforcement Learning* (RL), di mana *agent* belajar dengan cara berinteraksi langsung dengan lingkungannya untuk memaksimalkan *Reward* kumulatif (Ibrahim et al., 2024).

Algoritma RL klasik seperti *Deep* Q-Network (DQN) cukup berhasil pada permasalahan dengan ruang aksi diskrit. Namun, pada lingkungan dengan ruang aksi kontinu yang kompleks, algoritma tersebut menghadapi keterbatasan dari segi stabilitas dan kecepatan pembelajaran. Untuk mengatasi hal ini, dikembangkan algoritma berbasis *Policy* Gradient seperti *Proximal Policy Optimization* (*PPO*), yang dikenal lebih stabil dan efisien dalam mengoptimalkan kebijakan pada domain kontrol kontinu (V & Elumalai, 2025). Walaupun demikian, berbagai penelitian masih melaporkan bahwa *PPO* membutuhkan waktu pembelajaran yang lama serta berisiko menghasilkan kebijakan sub optimal karena eksplorasi yang kurang efektif (Ibrahim et al., 2024).

Kondisi ini menjadi celah penelitian yang penting untuk ditangani. Salah satu pendekatan yang banyak dikaji adalah *Reward Shaping*, yaitu teknik yang menambahkan sinyal *reward* tambahan agar *agent*t dapat lebih cepat menemukan jalur optimal. Pendekatan ini memungkinkan untuk mempercepat dan menstabilkan proses pembelajaran *agent*.

Selain itu, mayoritas penelitian RL masih menilai performa *agent* berdasarkan satu metrik saja, misalnya rata-rata *reward* (Larsen et al., 2021). Padahal, evaluasi tunggal sering kali tidak cukup menggambarkan kualitas *agent* secara menyeluruh. Oleh sebab itu, evaluasi multi-metrik — seperti *mean reward*, *success rate*, panjang episode, serta stabilitas pembelajaran — menjadi penting untuk memberikan penilaian yang lebih objektif dan komprehensif terhadap performa *agent* RL.

Berdasarkan permasalahan tersebut, penelitian ini membangun sebuah agen Deep Reinforcement Learning berbasis PPO yang dimodifikasi dengan teknik Reward Shaping pada simulasi robot Humanoid-v5. Secara konkret, penelitian ini merancang tiga komponen reward tambahan berbasis karakteristik fisik robot, yaitu stabilitas postur tubuh, efisiensi energi, dan ketahanan gerak yang ditambahkan di atas reward bawaan environment. Performa agen kemudian dibandingkan langsung dengan PPO Standar menggunakan lima metrik evaluasi: mean reward, success rate, episode length, torso stability, dan energy efficiency.

Berdasarkan uraian tersebut, permasalahan utama dalam penelitian ini adalah bagaimana meningkatkan stabilitas dan efisiensi pembelajaran *agent PPO* pada *Environment Humanoid* yang memiliki ruang aksi kontinu dan kompleks. Permasalahan ini muncul karena *PPO* cenderung mengalami fluktuasi *reward* yang tinggi dan lambat dalam mencapai konvergensi optimal, terutama pada lingkungan berdimensi tinggi. Untuk mengatasi permasalahan tersebut, penelitian ini mengusulkan penerapan teknik *reward Shaping* yang menambahkan sinyal *reward* berbasis domain agar *agent PPO* dapat belajar lebih cepat dan stabil. *Reward*

Shaping ini dirancang dengan mempertimbangkan beberapa komponen, seperti stabilitas postur tubuh, efisiensi energi, dan ketahanan gerakan, sehingga *agent* memiliki motivasi yang lebih seimbang antara menjaga keseimbangan dan bergerak maju secara efisien.

Selain itu, penelitian ini juga menerapkan pendekatan evaluasi multi-metrik, yaitu *mean reward*, *success rate*, episode *length*, torso *stabillity* dan energi *efifiency*. Dengan demikian, evaluasi hasil menjadi lebih menyeluruh dan dapat menggambarkan kemampuan *agent* secara lebih realistis.

Hasil pelatihan *agent* dalam simulasi ini berpotensi untuk diterapkan pada robot fisik nyata melalui pendekatan sim-to-real transfer di mana model yang telah dilatih dalam simulasi MuJoCo dapat dipindahkan ke platform robot sesungguhnya. Pendekatan ini menjadikan penelitian berbasis simulasi seperti ini memiliki nilai praktis yang nyata bagi pengembangan sistem kontrol robot di dunia nyata.

Dalam penelitian ini, *Environment Humanoid* dipilih karena kompleksitasnya yang tinggi dan relevansinya sebagai simulasi fisika berbasis MuJoCo (*Multi-Joint dynamics with Contact*). Lingkungan ini dirancang untuk menguji kemampuan *agent Reinforcement Learning* dalam mengontrol robot *Humanoid* berkaki dua di ruang aksi kontinu berdimensi tinggi, dengan tujuan menjaga keseimbangan serta bergerak ke depan secara stabil (Environment, 2025). Kompleksitas tersebut mencakup dinamika gerakan yang realistis, kebutuhan koordinasi motorik yang presisi, serta interaksi gaya dan momentum yang menyerupai kondisi fisik nyata. Dengan karakteristik demikian, penelitian ini tidak hanya berfokus pada peningkatan performa algoritma *PPO* melalui penerapan

reward Shaping, tetapi juga diharapkan dapat memberikan kontribusi pada pengembangan robotika cerdas, simulasi pergerakan manusia, serta teknologi pembelajaran kecerdasan buatan di lingkungan akademik.

1.2 Perumusan Masalah

Permasalahan dalam penelitian ini berfokus pada bagaimana mengimplementasikan *agent Deep Reinforcement Learning* menggunakan *Proximal Policy Optimization* dengan *Reward Shaping* dan evaluasi multimetrik pada *Environment Humanoid*.

1.3 Batasan Masalah

Untuk memperjelas ruang lingkup penelitian dan agar pembahasan lebih terfokus, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

a. Lingkungan penelitian terbatas pada *Environment Humanoid* dari *Gymnasium / MuJoCo*.
b. Algoritma yang digunakan yaitu *Proximal Policy Optimization* (*PPO*).
c. *Reward Shaping* hanya dilakukan dengan penambahan *Shaping* berbasis stabilitas dan gerakan.

1.4 Tujuan dan Manfaat

a. Tujuan

   Tujuan dari penelitian ini adalah untuk mengimplementasikan *agent Deep Reinforcement Learning* dengan menggunakan algoritma *Proximal Policy Optimization* dan *Reward Shaping* serta evaluasi multi-metrik pada *Environment Humanoid.*

b. Manfaat

a. Bagi Penulis

Penelitian ini bermanfaat untuk menambah pemahaman dan pengalaman praktis dalam mengimplementasikan algoritma *Deep Reinforcement* Learning, khususnya *PPO* dengan *Reward Shaping*, serta melatih kemampuan analisis dalam mengevaluasi performa *agent* menggunakan pendekatan multi-metrik.

b. Bagi STMIK Palangkaraya

Penelitian ini dapat memperkaya *khasanah* penelitian di bidang kecerdasan buatan, khususnya *Reinforcement Learning*, sehingga dapat menjadi referensi dan rujukan untuk penelitian selanjutnya bagi mahasiswa maupun dosen di lingkungan STMIK Palangkaraya.

c. Bagi Pengguna atau Aplikasi Praktis

Hasil penelitian ini dapat dimanfaatkan sebagai acuan dalam pengembangan sistem kontrol berbasis *agent* pada simulasi *Humanoid* maupun aplikasi robotika nyata, sehingga dapat mendukung kemajuan teknologi di bidang robotika, simulasi, dan kecerdasan buatan.

1.5 Sistematika Penulisan

Adapun sistematika penulisan dalam penulisan tugas akhir ini terdiri dari beberapa bab dan masing-masing bab membahas dan menguraikan pokok permasalahan yang berbeda, sebagai gambaran di sini penulis menyertakan garis-garis besarnya yaitu :

BAB I PENDAHULUAN

Bab ini berisikan latar belakang masalah, rumusan masalah, Batasan masalah, tujuan dan manfaat, serta sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisikan tentang tinjauan pustaka yang diambil dari penelitian yang relevan beserta susunan kajian teori yang disesuaikan dengan tema Tugas Akhir.

BAB III METODOLOGI PENELITIAN

Bab ini berisikan tentang tahapan yang dilakukan peneliti dalam mengumpulkan informasi atau data yang dibutuhkan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisikan tentang hasil dari penelitian dan pembahasan dari hasil yang didapat.

BAB V KESIMPULAN DAN SARAN

Bab ini berisikan tentang kesimpulan dan saran yang diberikan untuk pengembangan selanjutnya.

BAB II

TINJAUAN PUSTAKA

2.1 Kajian Teori

2.1.1 Konsep Dasar *Reinforcement Learning* (RL)

Reinforcement Learning (RL) adalah paradigma pembelajaran mesin di mana *agent* belajar mengambil keputusan melalui interaksi *trial-and-error* dengan lingkungan untuk memaksimalkan akumulasi imbalan (*Reward*) (Ghasemi & Ebrahimi, 2024; Murphy, 2025).

Komponen utama RL meliputi *agent*, lingkungan, fungsi kebijakan (*Policy*), fungsi nilai (*value function*), dan sinyal imbalan (*Reward signa*l) (Ghasemi & Ebrahimi, 2024; Murphy, 2025).

Proses *trial-and-error* menjelaskan bagaimana *agent* mengeksplorasi aksiaksi yang mungkin dan mengeksploitasi pengetahuan yang telah diperoleh, yang sering dikendalikan oleh strategi ε-*greedy* atau *softmax* (Zangirolami & Borrotti, 2024)

Fungsi nilai negara $\boldsymbol{V^{\pi}(s)}$ dan fungsi aksi $\boldsymbol{Q^{\pi}(s, a)}$ didefinisikan sebagai ekspektasi akumulasi imbalan mendatang di bawah kebijakan $\boldsymbol{\pi}$ (Ghasemi & Ebrahimi, 2024; Murphy, 2025):

$$V^{\pi}(\mathrm{s}) \;=\; \mathbb{E}_{\pi}[\textstyle\sum_{t=0}^{\infty} \gamma^{t} r_t \mid s_0 = \mathrm{s}] \qquad (1)$$

Secara konseptual, RL berbeda dengan *supervised learning* karena tidak ada label *Output* eksplisit, dan berbeda dengan *unsupervised learning* karena ada sinyal imbalan yang membimbing pembelajaran (Ghasemi & Ebrahimi, 2024; Murphy, 2025).

Penetapan sinyal imbalan yang tepat sangat menentukan konvergensi dan stabilitas *agent* RL (Ibrahim et al., 2024b). Konsep dasar ini menjadi landasan bagi implementasi *PPO* dengan *Reward Shaping* pada lingkungan *Humanoid*, karena pemahaman komponen dan fungsi nilai memengaruhi desain algoritma dan evaluasi *multi*-metrik (Murphy, 2025).

2.1.2 Markov Decision Process (MDP)

Markov Decision Process (MDP) adalah kerangka formal untuk memodelkan masalah RL, didefinisikan sebagai *tupel* ($\boldsymbol{S}$, $\boldsymbol{A}$, $\boldsymbol{P}$, $\boldsymbol{R}$, $\boldsymbol{\gamma}$) di mana $\boldsymbol{S}$ adalah himpunan status, A himpunan aksi, $\boldsymbol{P}(\boldsymbol{s} \mid \boldsymbol{s}, \boldsymbol{a})$ probabilitas transisi, $\boldsymbol{R}(\boldsymbol{s}, \boldsymbol{a})$ fungsi imbalan, dan $\boldsymbol{\gamma}$ adalah faktor diskonto yang mengatur seberapa besar pengaruh imbalan masa depan terhadap keputusan saat ini (Asadi et al., 2025). Sifat *Markov* memastikan bahwa distribusi status masa depan hanya bergantung pada status dan aksi saat ini, bukan pada sejarah lengkap, sehingga memungkinkan pembentukan model keputusan yang efisien dan terukur.

Persamaan Bellman untuk fungsi nilai taktis dirumuskan pada persamaan 2 dan 3.

$$V^{\pi}(s) = \sum_{s \in A} \mu\,(a|s) \sum_{s \epsilon \delta} P(s'|s,a)[R(s,a) = \gamma V^{\pi}(s')] \qquad (2)$$

Demikian pula, persamaan Bellman optimal untuk fungsi aksi Q*(s,a) menyatakan:

$$Q*(s,a) \;=\; \sum_{s'} P(s'|s,a)\;[R(s,a)[R(s,a) \;+\; \gamma max Q*(s',a')] \qquad (3)$$

MDP menyediakan landasan matematis bagi algoritma RL untuk mempelajari kebijakan optimal (Meggendorfer et al., 2025). Kejelasan definisi dan komponen MDP memastikan kerangka kerja yang konsisten dalam mengintegrasikan *PPO* dengan *reward shaping* pada permasalahan kontrol

Humanoid, di mana sifat stokastik dan dimensi status yang tinggi menuntut formulasi MDP yang matang.

2.1.3 Deep Reinforcement Learning (DRL)

Deep Reinforcement Learning (*DRL*) merupakan integrasi antara pembelajaran mendalam (*Deep* Learning) dan *Reinforcement Learning* (RL), di mana *Deep* neural networks (DNN) digunakan untuk mengaproksimasi fungsi nilai, kebijakan, atau model lingkungan dalam ruang keadaan dan aksi berdimensi tinggi (Sasso et al., 2023; Terven, 2025).

DRL memungkinkan *agent* belajar langsung dari data sensori yang kompleks tanpa rekayasa fitur manual, memberikan kemampuan adaptasi yang sangat baik di lingkungan dinamis (Terven, 2025)

Arsitektur populer dalam *DRL* adalah *actor-critic* yang memisahkan pengkodean kebijakan dan nilai, sehingga efektif untuk masalah kontrol kontinu seperti robotika *Humanoid*, aplikasinya meliputi berbagai bidang mulai dari navigasi robot, permainan cerdas, hingga diagnosis medis berbasis citra

Tantangan utama dalam *DRL* mencakup efisiensi sampel yang rendah, stabilitas pelatihan, dan eksplorasi optimal, yang memerlukan teknik seperti *replay buffer*, regularisasi, dan *tuning hyperparameter* yang cermat (Kim et al., 2023). f

2.1.4 Proximal Policy Optimization (PPO)

Proximal Policy Optimization (*PPO*) dikembangkan untuk mengurangi kompleksitas dan meningkatkan stabilitas pembaruan kebijakan dibanding Trust Region *Policy Optimization* (TRPO) (Z. Zhang et al., 2024). Fungsi objektif *PPO* dengan *clipping* dirumuskan sebagai:

$$L^{CLIP}(\theta) = \mathbb{E}[\min(r_t(\theta)\, A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)] \quad (4)$$

Dengan rasio probabilitas update :

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \quad (5)$$

Dan estimasi *advantage* A (Schulman et al., 2021). *PPO* terbukti lebih sederhana dan efisien daripada TRPO (L. Zhang et al., 2022) dan memiliki performa kompetitif dengan DDPG pada tugas kontinuitas aksi.

Perkembangan terkini menunjukkan adanya berbagai modifikasi *PPO*, seperti adaptasi *clipping* dinamis dan pendekatan berbasis distribusi Gaussian campuran untuk meningkatkan eksplorasi. Dengan keunggulan stabilitas dan kemudahan implementasi, *PPO* dipilih sebagai algoritma utama dalam penelitian ini yang memanfaatkan *Reward Shaping* dan evaluasi multi-metrik pada *Environment Humanoid*

2.1.4 Reward Shaping

Potential-Based Reward Shaping (PBRS) memperkenalkan fungsi *Shaping* $\Phi(s)$ untuk memodifikasi imbalan asli $\boldsymbol{R}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$ menjadi:

$$R(s, a, s') = R(s, a, s') + \gamma\Phi(s') - \Phi(s) \quad (6)$$

yang dijamin tidak mengubah kebijakan optimal, Metode lanjutan termasuk adaptasi otomatis Φ berbasis kurva pembelajaran, serta teknik *Reward Shaping* berbantuan kurva nilai (Jeon et al., 2023). Penelitian terbaru menunjukkan bahwa *Reward Shaping* adaptif dapat mempercepat konvergensi dan meningkatkan performa *agent* pada tugas kompleks, seperti berjalan pada *environtmet Humanoid* Penerapan *Reward Shaping* sangat relevan dalam penelitian ini untuk meningkatkan efisiensi *PPO* pada kontrol *Humanoid*.

2.1.5 Evaluasi Multi-Metrik

Evaluasi performa *agent* dalam *Reinforcement Learning* (RL) tidak cukup hanya dilihat dari rata-rata akumulasi *Reward*. Penilaian semacam itu cenderung terlalu sempit dan berisiko menutupi kelemahan model, terutama dalam konteks pelatihan jangka panjang atau ketika lingkungan tidak sepenuhnya deterministik. Karena itu, pendekatan multi-metrik menjadi semakin penting untuk memberikan gambaran yang lebih menyeluruh. Beberapa metrik tambahan yang kini banyak digunakan antara lain cumulative regret (Ji & Li, 2023), variance performa dan convergence rate (Wang et al., 2024), serta *robustness* terhadap gangguan eksternal seperti noise atau perturbasi aksi (Ayabe et al., 2025).

Penggunaan banyak metrik ini bukan hanya sekadar pelengkap, melainkan landasan penting dalam analisis perilaku *agent* yang realistis. Dalam lingkungan RL yang kompleks seperti *Humanoid locomotion*, metrik seperti *robustness* menjadi krusial karena sistem harus tetap stabil bahkan saat menghadapi *input* tak terduga atau fluktuasi kondisi (Zouitine et al., 2024). Studi terbaru juga menunjukkan bahwa pemilihan *hyperparameter* menjadi lebih efisien ketika didasarkan pada kombinasi metrik, dibanding hanya fokus pada *Reward* semata (Liu et al., 2025). Oleh karena itu, evaluasi multi-metrik sangat diperlukan dalam penelitian ini untuk menilai dampak dari *Reward Shaping* pada *agent PPO* secara lebih objektif dan holistik.

Dalam penelitian ini, evaluasi multi-metrik digunakan untuk menilai kinerja *agent PPO* dengan *Reward Shaping* secara lebih komprehensif. Lima metrik utama yang digunakan yaitu *mean reward*, *success rate*, episode *length*, stabilitas torso,

dan efisiensi energi, masing-masing merepresentasikan aspek kualitas kebijakan, keberhasilan tugas, efisiensi pelatihan, keseimbangan postur, dan penggunaan energi. Melalui kombinasi kelima metrik tersebut, penelitian ini menilai performa *agent* tidak hanya dari perolehan *reward*, tetapi juga dari stabilitas, efisiensi, dan realisme gerakan, sehingga hasil evaluasi menjadi lebih objektif dan menggambarkan secara mendalam pengaruh penerapan *reward Shaping*.

2.1.6 Environment Humanoid

Lingkungan *Humanoid* dalam simulasi fisika MuJoCo merupakan *benchmark* umum dalam penelitian *Reinforcement Learning* karena menghadirkan tantangan kontrol kontinu berdimensi tinggi. *Observation space*-nya terdiri dari sekitar 376 dimensi yang mencakup posisi, kecepatan, serta atribut fisik tubuh, sementara *action space*-nya berjumlah 17 dimensi yang merepresentasikan torsi pada sendi-sendi utama tubuh *robotic*. Tantangan utama dalam *Environment* ini pada gerakan, serta ketahanan terhadap gangguan *sensorik*. Struktur *Reward* umumnya mencakup komponen kecepatan, penalti energi, dan penalti jatuh, sebagaimana diterapkan dalam berbagai studi kontrol *locomotion*. Studi terbaru juga menunjukkan bahwa modifikasi skema *Reward* melalui *Reward Shaping* dapat meningkatkan efektivitas pembelajaran *agent* dalam lingkungan *Humanoid* yang kompleks (Jeon et al., 2023) Karena kompleksitas dan kemiripannya dengan skenario robotika nyata, *Environment* ini ideal untuk menguji algoritma seperti *Proximal Policy Optimization* (*PPO*) secara komprehensif.

2.1.7 Stable-*Baseline* s3 (SB3)

Stable-*Baseline* s3 (SB3) merupakan *framework open-source* berbasis Python yang menyediakan implementasi berbagai algoritma *Deep Reinforcement Learning* seperti *PPO*, *A2C*, *DQN*, *SAC*, dan *TD3*. *Framework* ini dikembangkan untuk menyediakan implementasi algoritma RL yang konsisten dan dapat digunakan sebagai *baseline* eksperimen yang dapat direproduksi. Selain itu, SB3 juga menyediakan fitur pendukung seperti *logging* TensorBoard, *vectorized environments*, serta *callback training* untuk mempermudah proses eksperimen *Reinforcement Learning* (Raffin et al., 2021).

2.1.7 Metode *Research and Development* (R&D)

Metode *Research and Development* (R&D) merupakan pendekatan penelitian yang bertujuan untuk menghasilkan suatu produk sekaligus menguji keefektifannya. Dalam penerapannya, proses R&D diawali dengan analisis kebutuhan sebagai dasar pengembangan produk, kemudian dilanjutkan dengan tahapan pengujian untuk menilai efektivitas produk tersebut (Fayrus et al., 2022). Tahapan R&D secara umum meliputi:

a *Research and Information Collecting*

Pada tahap awal dilakukan penelitian pendahuluan melalui studi pustaka, pengamatan langsung, serta identifikasi permasalahan yang relevan. Hasilnya dirangkum untuk menjadi dasar pengembangan produk.

b *Planning*

Tahap ini mencakup identifikasi dan definisi keterampilan yang dibutuhkan, perumusan tujuan penelitian, serta penyusunan rencana uji coba dalam skala kecil, termasuk validasi melalui *expert judgement*.

c *Develop Preliminary Form of Product*

Menyusun rancangan awal produk, yang dapat berupa materi pembelajaran, panduan penggunaan, maupun perangkat evaluasi, sehingga menghasilkan bentuk produk awal yang siap diuji.

d *Preliminary Field Testing*

Tahap ini merupakan pengujian awal terhadap produk dalam skala terbatas. Uji coba dilakukan untuk melihat kelayakan dasar, mendeteksi kekurangan, serta mengumpulkan data awal terkait kinerja produk. Hasil dari tahap ini menjadi bahan pertimbangan untuk perbaikan lebih lanjut.

e *Main Product Revision*

Produk direvisi berdasarkan hasil uji coba awal serta masukan dari pengguna dan pakar.

f *Main Field Testing*

Pada tahap ini produk diuji dalam skala yang lebih luas dengan kondisi penggunaan yang lebih bervariasi. Tujuannya adalah untuk memperoleh data yang lebih komprehensif mengenai efektivitas, stabilitas, dan keandalan produk sebelum memasuki tahap penyempurnaan.

g *Operational Product Revision*

Produk kembali diperbaiki dengan mempertimbangkan temuan serta masukan dari uji coba lapangan utama.

h *Operational Field Testing*

Produk final diuji dalam kondisi nyata untuk memastikan kelayakan dan efektivitasnya sebelum diterapkan secara luas.

i *Final Product Revision*

Produk disempurnakan berdasarkan hasil uji lapangan operasional sehingga diperoleh bentuk produk yang optimal.

j *Dissemination and Implementation*

Setelah produk akhir selesai dikembangkan dan melalui serangkaian uji efektivitas, tahap selanjutnya adalah diseminasi dan implementasi. Pada tahap ini, produk diperkenalkan dan diterapkan secara lebih luas kepada pengguna sasaran melalui kegiatan sosialisasi, pelatihan, maupun pendampingan. Selain itu, dilakukan juga pemantauan untuk memastikan produk dapat digunakan dengan baik sesuai tujuan serta memberikan manfaat yang optimal bagi pengguna.

2.2 Penelitian yang Relevan

Penulis menelusuri, menelaah dan mempelajari penelitian terdahulu yang relevan untuk menghindari duplikat dari pihak lain. Adapun hasil yang diperoleh dari penelusuran penelitian terdahulu sebagai berikut :

Tabel 2. 1 Penelitian Yang Relevan

No.	Penulis /Tahun	Topik Penelitian	Metode	Pembahasan	Hasil
1	Mingqi Yuan et al. (2023)	Automatic Intrinsic *Reward Shaping* (AIRS)	Pengaturan *Shaping Reward* intrinsik secara adaptif selaras fase pembelajaran	Memformulasikan pemilihan *Shaping Reward* sebagai masalah multi-armed bandit, di mana fungsi *Shaping* dipilih berdasarkan estimasi return saat Pelatihan.	AIRS berhasil meningkatkan eksplorasi dan performa dibanding *baseline* , dan mengurangi bias dalam kombinasi *Reward* eksternal & intrinsic.
2	Jeon, S. H., Heim, S. (2023)	Benchmarking Potential Based *Reward*s for *Learning Humanoid Locomotion*	Eksperimen *Deep* RL pada robot *Humanoid* dengan potential-based *Reward Shaping* (PBRS)	Membandingkan berbagai bentuk *Shaping Reward* (PBRS vs *Reward* biasa) dalam sistem *Humanoid* berdimensi tinggi; meneliti apakah PBRS efektif meningkatkan kecepatan konvergensi.	Ditemukan bahwa PBRS hanya memberikan peningkatan marginal pada kecepatan konvergensi dalam sistem berdimensi tinggi, tetapi lebih robust terhadap skala dibanding *Shaping* biasa

No.	Penulis /Tahun	Topik Penelitian	Metode	Pembahasan	Hasil
3	W. Chen et al. (2023)	*Static Standing Balance* dengan *PPO* + *Reward Shaping*	*PPO* yang dimodifikasi dengan *Reward Shaping* untuk tugas keseimbangan statis dalam model muskuloskeletal	Penelitian memodifikasi fungsi *Reward* untuk tugas balance (robot *Humanoid* atau model manusia) agar *agent* mampu mempertahankan pose statis lebih lama	Metode *PPO + Reward Shaping* berhasil menjaga keseimbangan lebih stabil dibanding *PPO* Standar tanpa *Shaping*
4	S. Chu et al. (2025)	Adaptive *Reward Shaping* for *DRL* (Adaptive *Reward* Shape *PPO* – ARS*PPO*)	Mengusulkan ARS*PPO* yaitu *PPO* dengan adaptasi *Reward Shaping* yang disesuaikan secara dinamis selama pelatihan.	Mengusulkan ARS*PPO* yaitu *PPO* dengan adaptasi *Reward Shaping* yang disesuaikan secara dinamis selama pelatihan	ARS*PPO* menunjukkan peningkatan efisiensi pelatihan dan performa dibanding *PPO* konvensional.
5	L. Berducci et al. (2025)	HPRS: Hierarchical Potential-based *Reward Shaping*	*Reward Shaping* otomatis berbasis spesifikasi tugas bertingkat (hierarchical formal requirements)	Menyajikan kerangka auto-*Shaping* yang menghasilkan *Reward Shaping* secara otomatis berdasarkan spesifikasi tugas, diuji pada berbagai simulasi termasuk *Humanoid locomotion* .	HPRS menunjukkan efektivitas dalam menghasilkan *Shaping Reward* tanpa intervensi manual, mempercepat pembelajaran di berbagai tugas robotik

Penelitian sebelumnya menunjukkan bahwa *Reward Shaping* efektif mempercepat konvergensi dan meningkatkan stabilitas pembelajaran pada *agent Reinforcement* Learning. Namun, sebagian besar studi masih berfokus pada satu metrik evaluasi dan satu aspek perilaku, serta jarang diterapkan pada *Environment* kompleks seperti *Humanoid*.

Berdasarkan analisis tersebut, terdapat celah penelitian berupa belum adanya integrasi antara *reward Shaping* multi-komponen dan evaluasi multimetrik pada *Environment Humanoid*. Penelitian ini merancang *reward Shaping* yang mencakup stabilitas, efisiensi energi, dan ketahanan gerak, serta mengevaluasinya secara multi-metrik untuk menghasilkan *agent PPO* yang lebih stabil dan efisien.

BAB III

METODOLOGI PENELITIAN

3.1 Tinjauan Umum

Penelitian ini bertujuan untuk mengembangkan *agent Deep Reinforcement Learning* (*DRL*) dengan menggunakan algoritma *Proximal Policy Optimization* (*PPO*) agar mampu belajar secara efektif dalam menyelesaikan permasalahan kontrol kontinu pada *Environment Humanoid*. Permasalahan yang menjadi fokus penelitian ini adalah proses pembelajaran *agent* yang cenderung lambat dalam mencapai konvergensi, kurang stabil selama proses pelatihan, serta penggunaan metrik evaluasi yang masih terbatas pada beberapa penelitian sebelumnya.

Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan teknik *Reward Shaping* guna memberikan sinyal umpan balik tambahan selama proses pembelajaran berlangsung. Selain itu, digunakan pula pendekatan evaluasi berbasis multi-metrik agar performa *agent* dapat dianalisis secara lebih menyeluruh. Dalam proses pelatihannya, *agent* akan berinteraksi secara langsung dengan *Environment* dengan tujuan memaksimalkan *reward* kumulatif sekaligus memperbaiki kebijakan (policy) berdasarkan pengalaman yang diperoleh selama proses pembelajaran.

Pada tahap implementasi, algoritma *PPO* dijalankan dengan menggunakan *framework* Stable-*Baseline* s3 (SB3), yaitu library *open-source* berbasis Python dan PyTorch yang dirancang untuk pengembangan algoritma *Deep Reinforcement* Learning. *Framework* ini menyediakan implementasi algoritma RL yang stabil serta dapat terintegrasi dengan *Environment Gymnasium* dan MuJoCo, sehingga

proses pelatihan, evaluasi, dan pengelolaan eksperimen *Reinforcement Learning* dapat dilakukan dengan lebih mudah dan terstruktur.

Penelitian ini meliputi beberapa tahapan, mulai dari perancangan arsitektur *agent*, implementasi algoritma *PPO* yang telah dimodifikasi dengan teknik *reward Shaping*, pengujian performa *agent* pada *Environment Humanoid*, hingga analisis hasil pelatihan menggunakan beberapa indikator evaluasi. Melalui penelitian ini diharapkan dapat diperoleh metode pembelajaran *Reinforcement Learning* yang lebih stabil, efisien, serta mampu beradaptasi dengan baik pada permasalahan kontrol yang kompleks.

3.2 Jenis Penelitian

Jenis penelitian yang dipilih adalah penelitian eksperimen terapan dengan karakteristik *Research and Development* (R&D). Penelitian eksperimen ini fokus pada pengembangan dan pengujian algoritma *PPO* yang dimodifikasi dengan teknik *Reward Shaping* serta menggunakan evaluasi multi-metrik untuk mengukur kinerja. Pemilihan metode R&D relevan karena penelitian ini tidak hanya menguji teori yang sudah ada, tetapi juga mengembangkan metode baru dan mengevaluasi keefektifannya pada permasalahan nyata yaitu kontrol *Humanoid*. Pada konteks ini, *agent* RL dikembangkan dan dioptimasi melalui proses eksperimen yang melibatkan pelatihan berulang dan pengujian hasil pada simulasi, sehingga pendekatan R&D dapat menyesuaikan dengan kebutuhan pengembangan sistem yang adaptif dan inovatif.

a. *Research and Information Collecting*

Tahap awal penelitian dilakukan dengan pengumpulan informasi melalui studi literatur mengenai *Reinforcement* Learning, algoritma *Proximal Policy Optimization* (*PPO*), teknik *reward Shaping*, serta praktik evaluasi multimetrik. Selain itu, peneliti melakukan observasi awal terhadap karakteristik *Environment Humanoid* berbasis MuJoCo untuk memahami ruang aksi, ruang keadaan, dan struktur *reward* bawaan sebagai dasar perancangan model. Pada tahap ini juga dilakukan kajian literatur mengenai *framework Deep Reinforcement Learning* yang digunakan dalam implementasi penelitian, yaitu Stable-*Baseline* s3 (SB3).

b. *Planning*

Tahap perencanaan menetapkan tujuan penelitian untuk menghasilkan *agent PPO* yang lebih stabil dan efisien melalui *reward Shaping*. Di sini ditentukan variabel dan parameter penting seperti arsitektur actor–critic, *Learning* rate, clip range, discount factor, komponen-komponen *reward*, serta metrik evaluasi yang akan digunakan (*mean reward*, *success rate*, episode *length*, stabilitas torso, dan efisiensi energi).

c. Develop Preliminary Form of Product

Berdasarkan perencanaan, dikembangkan bentuk awal produk berupa *agent PPO* dasar tanpa modifikasi *reward*. Implementasi model dilakukan menggunakan bahasa pemrograman Python dengan *framework* Stable-*Baseline* s3 (SB3) yang berjalan di atas PyTorch, serta

diintegrasikan dengan *Environment Gymnasium* dan MuJoCo. Pada tahap ini ditetapkan struktur jaringan saraf, mekanisme *logging* hasil pelatihan, serta prosedur pelatihan awal guna memastikan *agent* dapat berinteraksi dengan *Environment* dan menghasilkan kinerja dasar yang valid sebagai pembanding.

d. Preliminary Field Testing

Produk awal kemudian diuji secara terbatas untuk memperoleh gambaran awal performa. Pelatihan *PPO* standar dijalankan menggunakan stable-*baseline* s3 (SB3) pada *Humanoid* dengan jumlah langkah interaksi yang terkendali, sambil mengumpulkan metrik awal seperti total/*mean reward*, *success rate*, dan episode *length*. Analisis terhadap hasil ini digunakan untuk mengungkap kelemahan *baseline* —misalnya fluktuasi *reward* yang tinggi, kecenderungan jatuh, atau penggunaan energi yang boros—yang selanjutnya menjadi dasar perumusan *reward Shaping*

e. *Main Product Revision*

Berdasarkan temuan uji awal, produk direvisi dengan merancang *reward Shaping* berbasis domain yang menekankan tiga komponen utama: stabilitas postur (melalui indikator torso), efisiensi energi, dan ketahanan gerak. Bobot antar komponen disetel agar tercapai keseimbangan antara kemampuan bergerak maju dan menjaga keseimbangan. Verifikasi fungsi *reward* dilakukan melalui pelatihan singkat dan pemeriksaan logs untuk memastikan sinyal *reward* memandu perilaku yang diharapkan.

f. *Main Field Testing*

Versi model yang telah direvisi kemudian diuji pada skala lebih luas. Pelatihan *agent PPO* dengan *reward Shaping* dijalankan dalam horizon langkah yang lebih panjang serta beberapa seed bila diperlukan dan performanya dievaluasi secara multi-metrik. Perbandingan menyeluruh antara *PPO* standar dan *PPO*-RS dilakukan untuk menilai dampak *reward Shaping* terhadap stabilitas, efisiensi, dan kemampuan menyelesaikan tugas.

g. *Operational Product Revision*

Hasil uji utama menjadi dasar penyempurnaan lanjutan. Peneliti melakukan penyesuaian bobot *reward* antarkomponen, *tuning hyperparameter* (seperti *Learning* rate, clip epsilon, dan entropy coefficient), serta perbaikan teknis pada kode atau prosedur pelatihan untuk meningkatkan konvergensi dan konsistensi. Siklus train–evaluate–refine diulang sampai indikator kinerja menunjukkan perbaikan yang stabil.

h. *Operational Field Testing*

Model yang telah dioptimalkan kemudian diuji dalam kondisi operasional yang lebih representatif. Pelatihan akhir dijalankan dengan konfigurasi terbaik pada beberapa skenario awal dan variasi noise untuk menguji *robustness*. Seluruh metrik—*mean reward*, *success rate*, episode *length*, stabilitas torso, dan efisiensi energi—dikumpulkan dan

divisualisasikan untuk memastikan performa yang konsisten dan dapat direplikasi.

i. *Final Product Revision*

Berdasarkan hasil pengujian operasional, dilakukan penyempurnaan terakhir sehingga diperoleh model *PPO* dengan *reward Shaping* yang stabil, efisien, dan andal pada *Environment Humanoid*. Pada tahap ini pula disusun ringkasan hasil dalam bentuk laporan evaluasi multimetrik yang merangkum Perbandingan terhadap *baseline* , analisis tren pelatihan, serta implikasi praktis dari rancangan *reward*.

j. Dissemination *and* Implementation

Tahap akhir berfokus pada kegiatan diseminasi dan implementasi hasil penelitian. Pada tahap ini, peneliti menyiapkan dokumentasi lengkap yang mencakup laporan penelitian, kode sumber (*source* code), serta panduan replikasi eksperimen agar hasil penelitian dapat digunakan dan dikembangkan oleh peneliti lain. Sebagai bentuk keterbukaan ilmiah (*open* science), produk penelitian ini akan diunggah ke repositori publik GitHub, sehingga dapat diakses secara luas oleh komunitas akademik dan pengembang.

3.3 Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimen, karena melibatkan penerapan algoritma dan pengamatan langsung terhadap pengaruh perlakuan tertentu terhadap hasil yang diperoleh. Pendekatan ini dipilih untuk menguji efektivitas penerapan *Reward Shaping* terhadap performa *agent Proximal Policy*

Optimization (*PPO*) pada *Environment Humanoid*. Dalam konteks ini, peneliti berperan sebagai pengendali variabel bebas, yaitu parameter *reward* dan komponen *Shaping* yang digunakan dalam pelatihan *agent*. Berdasarkan prinsip umum metode eksperimen (Akbar et al., n.d.), penelitian ini disesuaikan dengan konteks pengujian algoritma *Reinforcement Learning* melalui tahapan-tahapan berikut:

a. Perumusan Masalah dan Hipotesis

   Penelitian ini diawali dengan perumusan masalah mengenai bagaimana penerapan *Reward Shaping* dapat memengaruhi stabilitas dan efisiensi pembelajaran *agent PPO*. Hipotesis yang diajukan adalah bahwa *Reward Shaping* dapat mempercepat konvergensi dan meningkatkan stabilitas pembelajaran pada *Environment Humanoid*.

b. Perancangan Eksperimen

   Tahap ini mencakup perancangan arsitektur jaringan saraf *actor–critic* untuk *agent PPO*, penentuan *hyperparameter* (*Learning* rate, gamma, clip range, dan batch size), penentuan pengunaan *framework* stable-*baseline* s3 serta perancangan fungsi *Reward Shaping* berbasis tiga komponen utama: stabilitas, efisiensi energi, dan ketahanan gerak.

c. Pelaksanaan Eksperimen

   Eksperimen dilakukan dengan mengimplementasikan *agent PPO* menggunakan Python dengan memanfaatkan *framework* stable-*baseline* s3 yang berjalan di atas PyTorch, serta menjalankan pelatihan pada *Environment Humanoid* berbasis MuJoCo. Dua skenario dijalankan: (1) *PPO* standar tanpa *Reward Shaping*, dan (2) *PPO* dengan *Reward Shaping*.

d. Pengumpulan dan Analisis Data

Selama proses pelatihan, sistem merekam hasil interaksi *agent* dengan *Environment* berupa trajectory (s , a , r , s) . Proses pelatihan dan pencatatan data eksperimen dilakukan menggunakan fitur *logging* yang tersedia pada *framework* Stable-*Baseline* s3 (SB3). Data ini kemudian dianalisis menggunakan lima metrik performa: *mean reward*, *success rate*, episode *length*, stabilitas torso, dan efisiensi energi.

e. Penarikan Kesimpulan dan Evaluasi

Berdasarkan hasil analisis metrik, dilakukan interpretasi terhadap efektivitas penerapan *Reward Shaping* dibanding *PPO* standar. Evaluasi dilakukan untuk menilai keunggulan maupun keterbatasan pendekatan yang dikembangkan.

Dengan tahapan tersebut, desain penelitian ini memastikan bahwa seluruh proses eksperimen berlangsung secara sistematis, terukur, dan dapat dievaluasi berdasarkan data empiris hasil simulasi, sehingga hubungan sebab-akibat antara penerapan *Reward Shaping* dan peningkatan performa *agent PPO* dapat diidentifikasi secara objektif.

3.4 Instrumen Penelitian

Instrumen penelitian yang digunakan adalah studi pustaka dengan mengumpulkan informasi dari berbagai referensi ilmiah, seperti jurnal, artikel, prosiding, dan buku yang relevan dengan topik penelitian. Literatur yang dikaji meliputi teori tentang *Reinforcement* Learning, algoritma *Proximal Policy Optimization* (*PPO*), teknik *Reward Shaping*, pendekatan evaluasi multi-metrik,

framework stable-*baseline* s3, serta penelitian-penelitian terdahulu yang berhubungan dengan implementasi *agent* RL pada *Environment* simulasi, khususnya MuJoCo *Humanoid*.

3.5 Analisis Kebutuhan

3.5.1 Kebutuhan Perangkat Keras

Perangkat keras yang digunakan dalam penelitian ini dapat dilihat di Tabel 3.1.

Tabel 3. 1 Kebutuhan Perangkat Keras

No	Hardware	Keterangan
1	Gpu	NVIDIA GEFORCE RTX 4060 16GB.
2	Processor	AMD Ryzen 5 7500F 6-Core Processor (12 CPUs), ~3,7GHz
3	Memory	16384 MB
4	Display	Xiaomi G24i 180hz 1920 x 1080 (23.8 inches)

3.5.2 Kebutuhan Perangkat Lunak

Perangkat lunak yang digunakan dalam penelian ini dapat dilihat di Tabel 3.2.

Tabel 3. 2 Kebutuhan Perangkat Lunak

No	Software	Keterangan
1	Python	Version: 3.13.5
2	Visual Studio code	Version: 1.105.0
3	Sistem Operasi	Windows 11 Pro 64-bit
4	CUDA Toolkit	Version: 13.0
5	Stable-*Baseline* s3	*Framework Deep Reinforcement Learning* berbasis PyTorch

3.5.1 Analisis Data

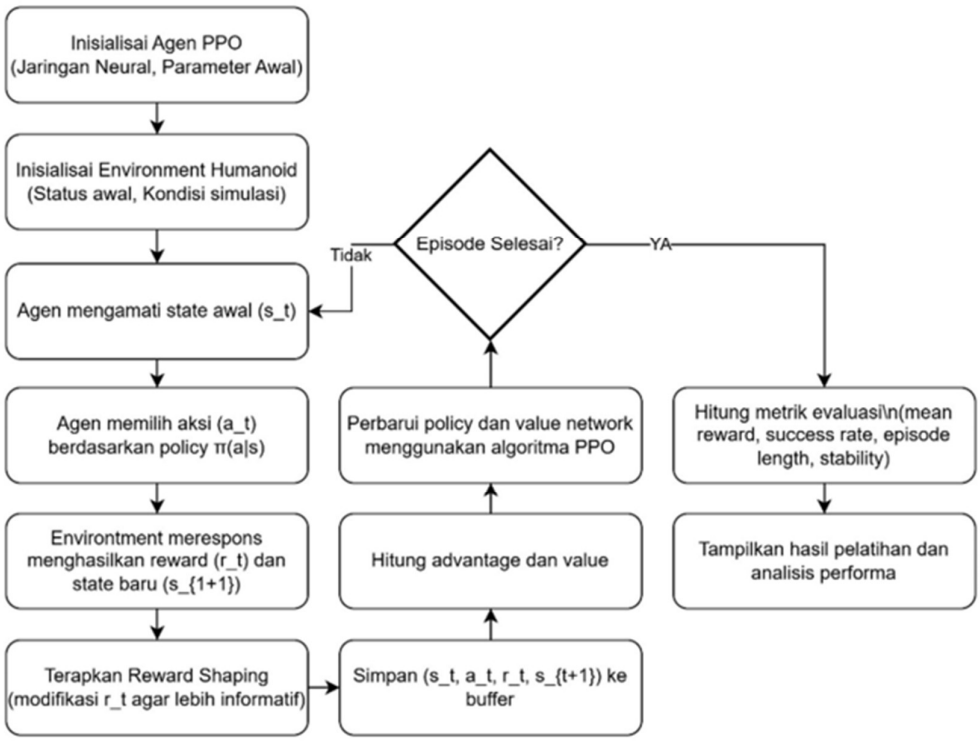
Data yang digunakan merupakan *trajectory* hasil interaksi *agent* dengan *Environment*, yang terdiri dari *state* (ss), *action* (aa), *Reward* (rr), dan *next state*

(s′s′). Data ini dikumpulkan secara otomatis selama proses pelatihan *agent* dalam *Environment* MuJoCo.

Setiap episode pelatihan menghasilkan data urutan interaksi antara *agent* dengan *Environment* yang digunakan sebagai dasar untuk mengupdate parameter model *PPO*. Dalam proses ini, teknik *reward Shaping* diterapkan untuk memodifikasi sinyal *reward* sehingga *agent* memperoleh umpan balik yang lebih informatif selama pembelajaran. Data hasil interaksi tersebut kemudian dianalisis berdasarkan performa *agent* menggunakan pendekatan evaluasi multi-metrik. Analisis ini bertujuan untuk menilai efektivitas penerapan *reward Shaping* dalam meningkatkan stabilitas pembelajaran serta kualitas kebijakan yang dihasilkan oleh *agent*.

3.5.2 Analisis Proses

Analisis proses dilakukan untuk menggambarkan alur kerja sistem secara menyeluruh, mulai dari inisialisasi *agent* hingga tahap evaluasi performa. Tujuannya adalah untuk memastikan setiap komponen penelitian—baik algoritma *PPO*, modul *reward Shaping*, maupun *Environment Humanoid*—berjalan sesuai rancangan dan saling berinteraksi secara terukur. adapun langkah-langkah yang dilakukan ditujukan pada Gambar 3.1 :



Gambar 3. 1 Flowchart Proses Pelatihan *Agent PPO* dengan *Reward Shaping*

Flowchart pada gambar 3.1 dijelaskan sebagai berikut:

a. Inisialisasi *Agent PPO*

Agent PPO diinisialisasi dengan parameter awal jaringan saraf *actor–critic* yang akan digunakan untuk menentukan kebijakan (*Policy*) dan memperkirakan nilai keadaan (*value*). Inisialisasi ini dilakukan menggunakan *framework* Stable-*Baseline* s3 (SB3) yang menyediakan implementasi algoritma *PPO* berbasis PyTorch untuk membangun model *actor–critic* secara otomatis.

b. *Inisialisasi Environment Humanoid*

Environment Humanoid diatur untuk menentukan status awal simulasi, termasuk posisi awal, orientasi tubuh, serta kondisi dinamika fisika dari

robot *Humanoid*. *Environment* ini diintegrasikan dengan *agent* melalui *interface Gymnasium* yang didukung oleh *framework* Stable-*Baseline* s3 (SB3).

c. *Agent* Mengamati State Awal

Agent mulai mengamati *state* awal (s) dari *Environment* untuk memperoleh informasi posisi, kecepatan sendi, dan orientasi tubuh sebagai dasar pengambilan keputusan pertama.

d. *Agent* Memilih Aksi

Berdasarkan *Policy* $\pi(a \mid s)$, *agent* memilih aksi (a) yang paling sesuai dengan kondisi *state* saat ini. Aksi ini berupa nilai kontinu yang mengontrol pergerakan sendi atau torsi pada robot *Humanoid*.

e. *Respon Environment*

Environment mengeksekusi aksi tersebut dan memberikan *feedback* berupa *reward* bawaan (r) serta *state* baru (s).

f. Penerapan *Reward Shaping*

Sinyal *reward* dari *Environment* dimodifikasi dengan menambahkan komponen tambahan seperti stabilitas torso, efisiensi energi, dan ketahanan postur. Proses ini menghasilkan *reward* total (r) yang lebih informatif bagi *agent* dalam belajar.

g. Penyimpanan *Trajectory*

Pada implementasi menggunakan Stable-*Baseline* s3 (SB3), data interaksi tersebut disimpan secara otomatis dalam *rollout buffer* yang digunakan oleh algoritma *PPO* selama proses pelatihan.

h. Hitung *Advantage* dan *Value*

Setelah sejumlah langkah tertentu, sistem menghitung *advantage* dan *value estimation* untuk setiap pasangan *state–action* menggunakan *Generalized Advantage Estimation* (GAE). Dalam implementasi menggunakan Stable-*Baseline* s3, proses perhitungan *advantage* dan *value estimation* ini dilakukan secara otomatis oleh modul *PPO* yang tersedia pada *framework* tersebut.

i. Pembaruan *Policy* dan *Value* Network

Parameter jaringan *actor* dan *critic* diperbarui menggunakan algoritma *PPO* dengan fungsi objektif *clipped surrogate*, untuk menjaga stabilitas dan mencegah perubahan kebijakan yang berlebihan. Proses pembaruan parameter jaringan ini dijalankan oleh algoritma *PPO* yang diimplementasikan dalam *framework* Stable-*Baseline* s3 (SB3) selama tahap *training* model.

j. Episode selesai

Setelah pembaruan parameter dilakukan, sistem akan memeriksa apakah episode pelatihan telah mencapai kondisi terminal atau batas langkah maksimum simulasi. Jika episode belum selesai (done = False), *agent* akan kembali ke langkah observasi *state* berikutnya (s)dan melanjutkan siklus interaksi dengan *Environment*. Hal ini memungkinkan *agent* untuk terus belajar secara bertahap melalui pengalaman berulang dalam episode yang sama. Namun jika episode

telah selesai (done = True), maka proses akan di lanjutkan ke tahapan berikutnya.

k. Perhitungan Metrik Evaluasi

Pada tahap ini dilakukan perhitungan terhadap beberapa metrik evaluasi utama, yaitu *mean reward*, *success rate*, episode *length*, stabilitas torso, dan efisiensi energi. Setiap metrik memberikan informasi berbeda tentang kemampuan *agent*, baik dari sisi efektivitas kebijakan, kestabilan gerak, maupun efisiensi dalam penggunaan energi. Hasil perhitungan metrik ini digunakan untuk menggambarkan performa *agent* setelah setiap episode pelatihan.

l. Penampilan Hasil Pelatihan dan Analisis Performa

Setelah metrik evaluasi dihitung, sistem menampilkan nilai-nilai hasil evaluasi tersebut dalam bentuk keluaran numerik atau grafik metrik untuk setiap episode. Tampilan ini berfungsi untuk memantau perkembangan performa *agent* selama proses pelatihan dan memastikan bahwa penerapan *Reward Shaping* memberikan pengaruh positif terhadap pembelajaran *agent PPO*.

Dengan cara ini, *agent* belajar secara bertahap dan evaluasi dilakukan untuk membandingkan dua skenario: *PPO* standar dan *PPO* dengan *Reward Shaping*.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Hasil

4.1.1 Implementasi PPO Standar (Baseline)

Penelitian ini diimplementasikan menggunakan bahasa pemrograman Python 3.13.5 dengan *framework* Stable-*Baseline* s3 (SB3) berbasis PyTorch. Simulasi fisika menggunakan MuJoCo melalui antarmuka *Gymnasium* dengan *Environment Humanoid*-v5. Dalam penelitian ini, diterapkan dua skenario pelatihan yang dipisah, yaitu *PPO* standar sebagai *baseline* dan *PPO* dengan metode *reward shaping*. Kedua skenario menggunakan *hyperparameter* yang identik agar Perbandingan yang dihasilkan adil dan valid secara ilmiah. Konfigurasi *hyperparameter* yang digunakan pada kedua skenario dapat dilihat pada Tabel 4.1.

Tabel 4. 1 Konfigurasi hyperparameter PPO

Parameter	Nilai	Keterangan
Environment	*Humanoid*-v5	*MuJoCo via Gymnasium*
Total *Timestep*s	50.000.000	Sama untuk kedua skenario
n_envs	18	Jumlah *Environment* paralel
n_steps	2.048	Langkah per *rollout*
batch_size	512	Ukuran *batch gradient update*
n_epochs	10	*Epoch* per *update*
learning_rate	0,0003	Linear *schedule*
gamma (γ)	0,99	Faktor diskonto
gae_lambda (λ)	0,95	GAE lambda
clip_range (ε)	0,2	*PPO clipping* parameter
seed	42	*Seed reprodusibilitas*

Pada skenario pertama, *agent PPO* Standar dilatih dengan menggunakan *reward* bawaan dari *Environment Humanoid*-v5 tanpa adanya modifikasi. *Reward* bawaan *Environment* terdiri dari komponen kecepatan gerak maju dan penalti apabila robot jatuh. Skenario ini digunakan sebagai *baseline* untuk mengukur performa dasar algoritma *PPO* sebelum diterapkan teknik *Reward Shaping*.

4.1.2 Implementasi PPO dengan Reward Shaping

Pada skenario kedua, *agent PPO* dilatih dengan menerapkan teknik *Reward Shaping*. Penelitian ini merancang komponen *Reward Shaping* secara spesifik sesuai karakteristik fisik robot *Humanoid*. *Reward* bawaan *Environment* dimodifikasi dengan menambahkan tiga komponen tambahan yang mencerminkan aspek penting dalam pengendalian gerak robot, yaitu stabilitas postur tubuh, efisiensi penggunaan energi, dan ketahanan bertahan. Adapun rumus *reward* total yang digunakan adalah sebagai berikut:

$$R_total = R_env + w1 \times stability + w2 \times energy + w3 \times endurance \quad (7)$$

Keterangan:

a. R_total = Total *reward* yang diterima *agent* pada setiap langkah waktu.
b. R_env = *Reward* bawaan *Environment Humanoid*-v5 (kecepatan maju + penalti jatuh).
c. w1 = Koefisien stabilitas postur.
d. *stability* = Komponen *reward* stabilitas torso (ketinggian tubuh + orientasi tegak).
e. w2 = Koefisien efisiensi energi .
f. *energy* = Penalti konsumsi energi berupa kuadrat torsi sendi (bernilai negatif).

g. w3 = Koefisien ketahanan / *survival* bonus.

h. *endurance* = Bonus flat per langkah selama *agent* masih berdiri dan tidak jatuh.

Nilai koefisien pada masing-masing komponen dipilih berdasarkan pada tingkat prioritas dalam proses pembelajaran *agent*. Komponen *stability* diberi bobot tertinggi yaitu 1,0 karena menjaga keseimbangan postur merupakan tantangan utama pada *Environment Humanoid*. Komponen *energy* diberi bobot 0,5 untuk mendorong gerakan yang halus tanpa terlalu membatasi eksplorasi *agent*. Sementara komponen *endurance* diberi bobot terkecil yaitu 0,1 sebagai dorongan tambahan agar *agent* tetap bertahan dan tidak mudah jatuh di setiap langkah. Penjelasan lebih rinci mengenai ketiga komponen tersebut dapat dilihat pada Tabel 4.2.

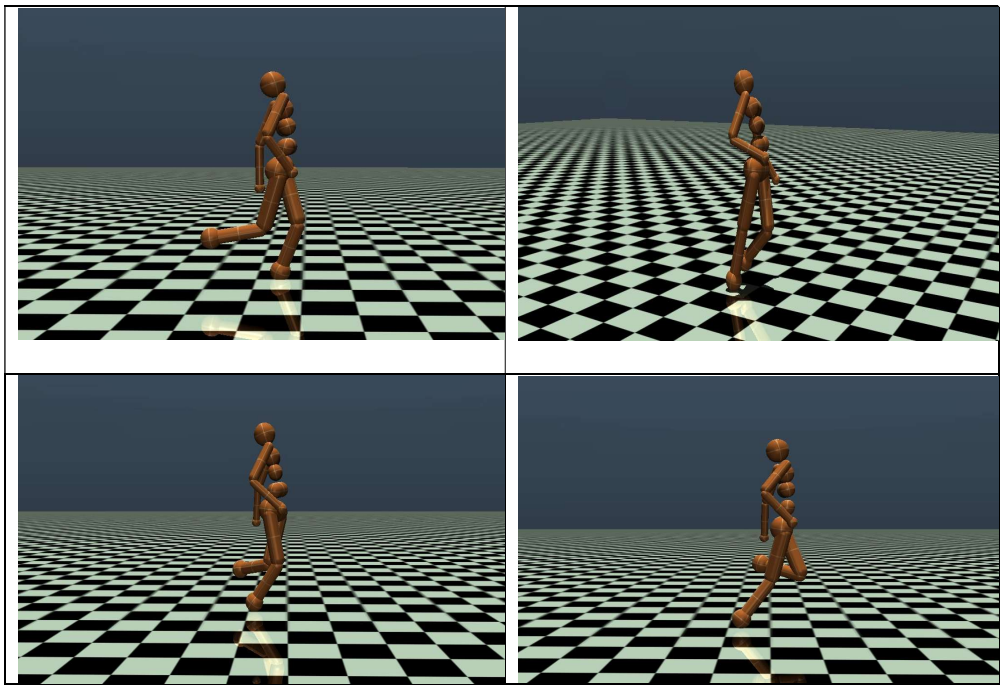
Tabel 4. 2 Komponen dan Koefisien *Reward Shaping*

Komponen	Variabel	Nilai	Fungsi
Stability (w1)	stability_coef	1,0	*Reward* stabilitas torso (ketinggian tubuh + orientasi tegak)
Energy (w2)	energy_coef	0,5	Penalti konsumsi energi berupa kuadrat torsi sendi
Endurance (w3)	endurance_coef	0,1	Bonus flat per langkah selama *agent* masih berdiri

Setelah kedua skenario selesai dikonfigurasi, proses pelatihan dijalankan menggunakan 18 *Environment* paralel (n_envs = 18) dengan total 50.000.000 *timestep*. Penggunaan *Environment* paralel bertujuan untuk mempercepat pengumpulan data pengalaman *agent*, sehingga proses konvergensi dapat dicapai dengan lebih efisien. Selama pelatihan berlangsung, sistem secara otomatis

mencatat data *trajektori* berupa *state*, *action*, *reward*, dan *next state* yang digunakan untuk memperbarui parameter jaringan *actor-critic* pada setiap siklus *update PPO*.

Gambar 4.1 menunjukkan hasil pelatihan *agent PPO* dengan penerapan *Reward Shaping* pada *Environment Humanoid*-v5. Dari gambar tersebut, terlihat bahwa robot *Humanoid* berhasil mengembangkan kemampuan berjalan dengan postur tubuh yang tegak dan langkah yang stabil, hal ini mengindikasikan efektivitas komponen *stability reward* dan *energy efficiency* yang diterapkan dalam proses pelatihan.



Gambar 4. 1Hasil Pelatihan *Agent PPO + Reward Shaping* pada *Environment Humanoid*-v5

4.2 Pengujian dan Perbandingan

4.2.1 Hasil PPO Standar (Baseline)

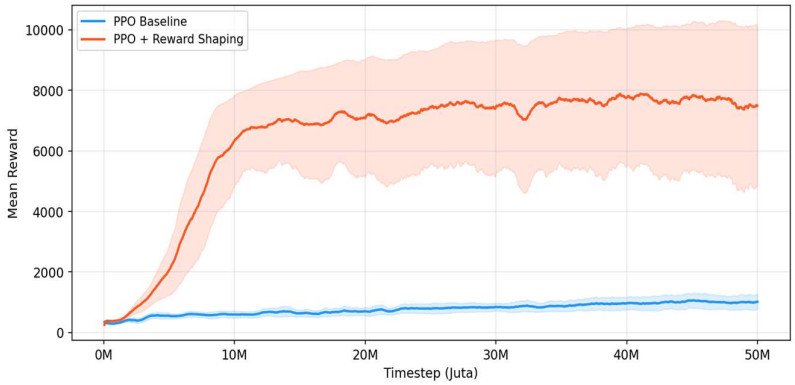
PPO Standar sebagai baseline menunjukkan performa yang terbatas selama 50.000.000 timestep pelatihan. Nilai mean reward tertinggi yang dicapai adalah 1.215,92 pada sekitar timestep ke-44,2 juta, namun cenderung stagnan setelahnya. Episode length rata-rata hanya mencapai 222,8 langkah, menandakan robot sering jatuh dalam waktu singkat. Metrik torso stability tidak dapat terukur, menunjukkan ketidakmampuan agen mempertahankan postur tegak secara konsisten. Success rate tercatat 0%, artinya agen tidak pernah berhasil menyelesaikan episode tanpa jatuh.

Pengujian dilakukan dengan menjalankan proses pelatihan pada kedua skenario hingga mencapai 50.000.000 *timestep* di *Environment Humanoid*-v5. Evaluasi dilakukan secara berkala setiap 50.000 *timestep* menggunakan 50 episode *deterministik* (*deterministic=true*) agar performa *agent* dapat diukur secara konsisten. Hasil pengujian disajikan dalam bentuk kurva pelatihan, tabel Perbandingan lima metrik utama, dan ringkasan hasil evaluasi akhir.

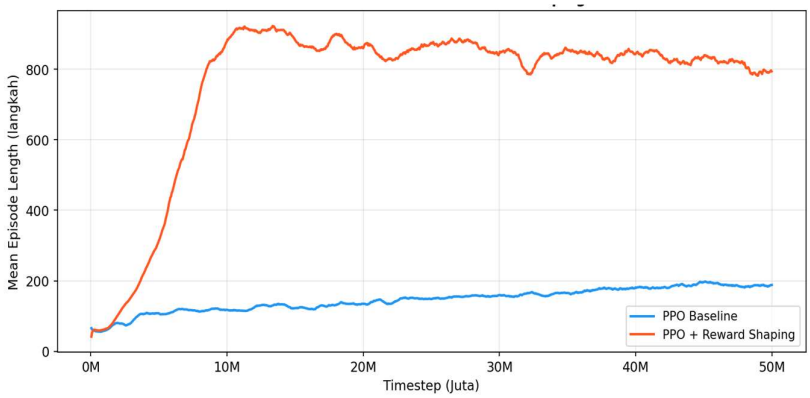
4.2.2 Hasil PPO dengan Reward Shaping

PPO dengan Reward Shaping menunjukkan peningkatan performa yang signifikan dibandingkan baseline. Nilai mean reward terbaik mencapai 8.641,51 pada timestep ke-38.889.108 lebih cepat dari baseline dan 611% lebih tinggi. Episode length mencapai puncak 934,96 langkah, jauh melampaui baseline 222,8 langkah. Torso stability berhasil terukur dengan nilai 1,363 berkat komponen stability reward (w1=1,0), dan efisiensi energi 46% lebih baik dibanding baseline.

Gambar 4. 2 Perbandingan Kurva *Mean Reward PPO Baseline* vs *PPO + Reward Shaping*



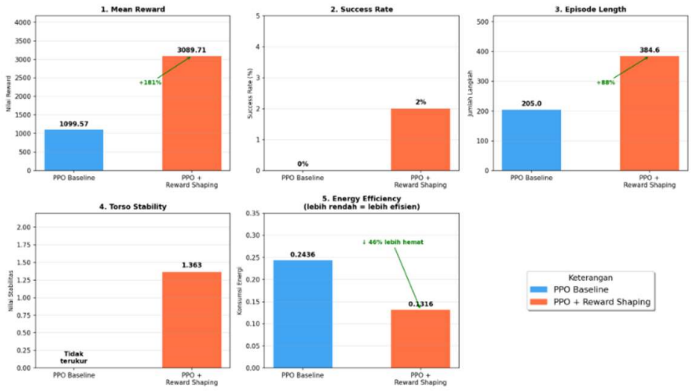
Berdasarkan Gambar 4.2, terdapat perbedaan yang signifikan antara kedua skenario. *PPO baseline* hanya mampu mencapai *mean reward* tertinggi sebesar 1.215,92 pada sekitar *timestep* ke-44,2 juta dan cenderung stagnan setelahnya. Sementara itu, *PPO* dengan *Reward Shaping* menunjukkan pertumbuhan yang jauh lebih pesat, mencapai *mean reward* tertinggi sebesar 8.641,51 pada *timestep* ke-38,9 juta dengan nilai akhir sebesar 7.233,82.



Gambar 4. 3 Perbandingan Kurva Episode *Length PPO Baseline* vs *PPO + Reward Shaping*

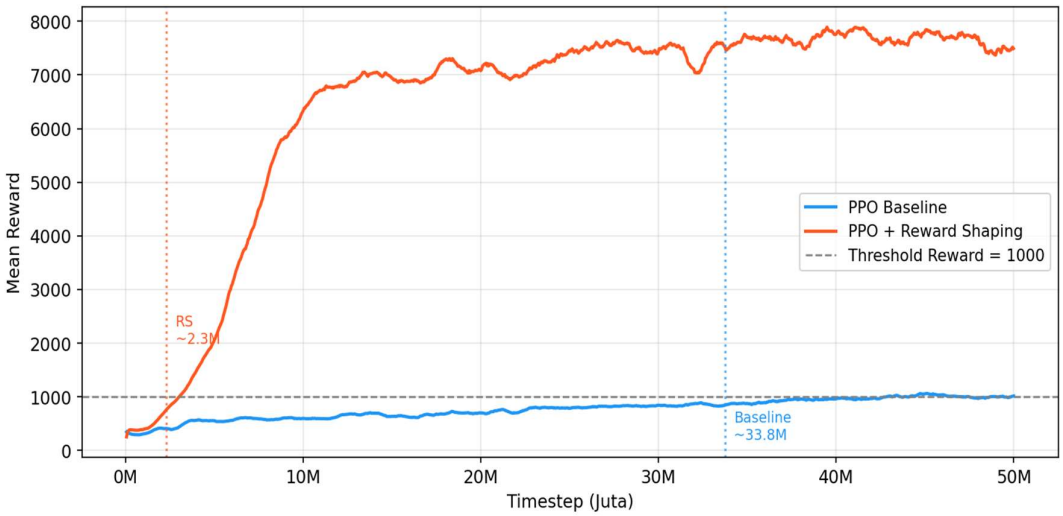
Gambar 4.3 menampilkan Perbandingan episode *length*, yaitu jumlah langkah yang berhasil dilakukan robot sebelum mengalami jatuh. *PPO baseline* hanya mencapai rata-rata episode *length* akhir sebesar 188 langkah, sedangkan *PPO*

dengan *reward shaping* mampu mencapai 767 langkah di akhir pelatihan dengan puncak tertinggi 934 langkah. Hal ini menunjukkan bahwa *agent* dengan *reward shaping* jauh lebih stabil dalam menjaga keseimbangan tubuh *humanoid*.



Gambar 4. 4 Perbandingan Nilai Akhir *Reward* dan Episode *Length* (Diagram Batang)

Gambar 4.4 menyajikan Perbandingan nilai metrik akhir secara visual dalam bentuk diagram batang. Terlihat bahwa seluruh metrik menunjukkan keunggulan *PPO* dengan *reward shaping* secara konsisten dan substansial dibandingkan *PPO Baseline* , dengan perbedaan yang cukup signifikan.



Gambar 4. 5 Perbandingan Kecepatan Konvergensi *PPO Baseline* vs *PPO + Reward Shaping*

Gambar 4.5 menampilkan Perbandingan kecepatan konvergensi kedua skenario menggunakan *threshold reward* 1.000 sebagai acuan. Hasilnya, *PPO* dengan *reward shaping* mencapai *threshold* tersebut jauh lebih awal dibandingkan *PPO baseline* , Hal ini menunjukan bahwa secara efektif mempercepat proses pembelajaran *agent*.

4.2.3 Perbandingan PPO Standar dan PPO + Reward Shaping

Tabel 4. 3 Perbandingan Lima Metrik Evaluasi Utama

Metrik	*PPO Baseline*	*PPO + Reward Shaping*
Mean Reward	1.099,57	3.089,71
Success rate	0,0 (0%)	0,02 (2%)
Episode *Length* (langkah)	205,0	384,6
Torso Stability	– (tidak terukur)	1,363
Energy Efficiency	0,2436	0,1316

Tabel 4.3 menyajikan Perbandingan lima metrik evaluasi utama. Berdasarkan hasil tersebut, *PPO* dengan *reward shaping* menunjukkan performa yang lebih baik dibandingkan *PPO baseline* pada seluruh metrik yang diukur. Nilai *mean reward* meningkat dari 1.099,57 menjadi 3.089,71, sementara episode *length* juga mengalami peningkatan dari 205,0 menjadi 384,6 langkah. *Success rate PPO baseline* tercatat 0% sedangkan *PPO* dengan *reward shaping* berhasil mencapai 2%, yang menunjukan bahwa *agent* mulai mampu menyelesaikan episode tanpa jatuh. Metrik torso *stability* hanya berhasil diukur pada skenario *reward shaping* dengan nilai 1,363, sementara pada *PPO baseline* nilai ini tidak dapat terukur

karena robot tidak berhasil mempertahankan postur tegak secara konsisten. Dari sisi efisiensi energi, *PPO* dengan *reward shaping* menunjukkan nilai yang lebih rendah (0,1316 vs 0,2436), yang berarti *agent* bergerak dengan konsumsi energi yang lebih hemat dan efisien.

Tabel 4. 4 Perkembangan Performa *Agent* Selama Proses Pelatihan

Metrik	***PPO Baseline***	***PPO + Reward Shaping***
Best Mean Reward	1.215,92	8.641,51
Final Mean Reward	1.017,53	7.233,82
Best Episode *Length* (langkah)	222,8	934,96
Final Episode *Length* (langkah)	188,0	767,02
Timestep saat Performa Terbaik	44.200.000	38.889.108
Peningkatan *Mean Reward* (%)	–	+611,0%
Peningkatan Episode *Length* (%)	–	+320,2%

Tabel 4.4 menampilkan perkembangan performa *agent* selama proses pelatihan 50.000.000 *timestep*, berdasarkan data kurva pelatihan (*Learning curve*) pada kedua skenario. Tabel ini memberikan gambaran sejauh mana kemampuan *agent* berkembang dari awal hingga akhir pelatihan, dilihat dari nilai performa terbaik yang pernah dicapai (*best*) dan nilai performa pada akhir pelatihan (*final*). Dari sisi perkembangan *mean reward*, *PPO* dengan *reward shaping* menunjukkan peningkatan yang jauh lebih signifikan, dengan nilai terbaik mencapai 8.641,51 dan nilai akhir sebesar 7.233,82. Sementara itu, *PPO baseline* hanya mampu mencapai nilai terbaik sebesar 1.215,92 dan nilai akhir 1.017,53, dengan selisih peningkatan mencapai 611,0%. Pola yang sama juga terlihat pada episode *length*, di mana *PPO*

dengan *reward shaping* mencapai puncak 934,96 langkah dan bertahan di 767,02 langkah pada akhir pelatihan, dibandingkan *PPO baseline* yang hanya mencapai 222,8 langkah pada puncaknya dan turun menjadi 188,0 langkah di akhir pelatihan, dengan peningkatan sebesar 320,2%.

Perbedaan dalam kecepatan pencapaian performa terbaik dapat di lihat pada *PPO* dengan *reward shaping* yang mencapai puncak performanya lebih awal, yaitu pada *timestep* ke-38.889.108, sementara *PPO baseline* baru mencapai puncaknya pada *timestep* ke-44.200.000. Ini menunjukkan bahwa *reward shaping* tidak hanya meningkatkan kualitas akhir performa *agent*, tetapi juga mempercepat proses perkembangan *agent* selama pelatihan berlangsung.

4.3 Pembahasan

Hasil eksperimen secara konsisten menunjukkan bahwa penerapan *reward shaping* pada algoritma *PPO* mampu meningkatkan performa secara signifikan pada *environment humanoid*-v5. Temuan ini mendukung hipotesis penelitian bahwa modifikasi sinyal *reward* berbasis domain dapat membantu mempercepat sekaligus menstabilkan proses pembelajaran *agent*, terutama pada lingkungan kontrol kontinu dengan dimensi yang tinggi.

4.3.1 Analisis Performa PPO Standar (Baseline)

PPO Standar terbukti mengalami keterbatasan nyata pada lingkungan Humanoid yang berdimensi tinggi. Dengan hanya mengandalkan reward bawaan berupa kecepatan maju dan penalti jatuh, agen tidak mendapatkan cukup sinyal untuk mempelajari perilaku kompleks seperti menjaga keseimbangan tubuh. Akibatnya, konvergensi tercapai terlambat (timestep ke-44.200.000) dan nilai akhir

yang dihasilkan jauh dari optimal — mengonfirmasi bahwa PPO standar tanpa modifikasi memang belum memadai untuk environment sekompleks Humanoid-v5.

4.3.2 Analisis Performa PPO dengan Reward Shaping

Dari sisi *mean reward*, *PPO* dengan *reward shaping* mencapai nilai terbaik sebesar 8.641,51, jauh lebih tinggi dibandingkan *PPO baseline* yang hanya mencapai 1.215,92, dengan peningkatan sebesar 611,0%. Hal ini menunjukkan bahwa penambahan komponen *reward* seperti stabilitas postur, efisiensi energi, dan survival bonus mampu memberikan sinyal pembelajaran yang lebih informatif. Dengan sinyal tersebut, *agent* tidak hanya belajar untuk bergerak maju, tetapi juga terdorong untuk menjaga keseimbangan serta bergerak secara lebih efisien.

Peningkatan yang signifikan juga terlihat pada metrik episode *length*. *PPO* dengan *reward shaping* mampu bertahan hingga rata-rata 767 langkah per episode di akhir pelatihan, sedangkan *PPO baseline* hanya mencapai 188 langkah, dengan peningkatan sebesar 320,2%. Hal ini mengindikasikan bahwa *agent* dengan *reward shaping* memiliki kebijakan yang lebih stabil dalam mempertahankan keseimbangan tubuh. Kondisi ini tidak lepas dari pengaruh komponen *stability reward* (w1 = 1,0) yang mendorong *agent* untuk menjaga posisi torso tetap berada pada ketinggian dan orientasi yang diharapkan.

Metrik torso *stability* menjadi indikator paling jelas terhadap efektivitas *reward shaping*. *PPO baseline* tidak mampu menghasilkan nilai yang terukur pada metrik ini, sedangkan *PPO* dengan *reward shaping* mencapai nilai 1,363. Hal ini menunjukkan bahwa tanpa adanya komponen khusus untuk stabilitas, *agent*

cenderung mengabaikan postur tubuh selama masih bisa memperoleh *reward* dari pergerakan. Sebaliknya, dengan adanya komponen tersebut, *agent* belajar menjadikan postur tegak sebagai bagian penting dari strateginya.

Dari sisi efisiensi energi, *PPO* dengan *reward shaping* menunjukkan performa yang lebih baik dengan nilai yang lebih rendah (0,1316 dibandingkan 0,2436), yang berarti penggunaan energi menjadi lebih hemat hingga 46,0%. Hal ini merupakan dampak dari penerapan *energy penalty* (w2 = 0,5) yang membatasi penggunaan torsi secara berlebihan. Akibatnya, *agent* cenderung mengembangkan pola gerakan yang lebih halus dan terkoordinasi, yang juga berkontribusi pada peningkatan stabilitas.

Selain itu, kecepatan konvergensi juga menunjukkan keunggulan pada skenario *reward shaping*. *PPO* dengan *reward shaping* mencapai performa terbaik pada *timestep* ke-38.889.108, lebih cepat dibandingkan *PPO baseline* yang mencapainya pada *timestep* ke-44.200.000. Temuan ini sejalan dengan teori *potential-based reward Shaping* (PBRS) yang menyatakan bahwa desain *reward* yang tepat dapat mempercepat konvergensi tanpa mengubah kebijakan optimal.

4.3.3 Perbandingan dan Implikasi

Meskipun demikian, nilai *success rate* pada kedua skenario masih tergolong rendah, yaitu 0% pada *PPO baseline* dan 2% pada *PPO* dengan *reward shaping*. Hal ini dapat dimaklumi karena *environment humanoid*-v5 termasuk salah satu yang paling kompleks dalam *benchmark MuJoCo*, dengan dimensi *observation* dan *action* yang sangat tinggi. Namun, peningkatan dari 0% menjadi 2% tetap

menunjukkan adanya progres, di mana *agent* mulai mampu menyelesaikan episode tanpa jatuh.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil implementasi dan pengujian yang telah dilakukan mengenai *agent deep reinforcement learning* menggunakan *proximal policy optimization* dengan *reward shaping* dan evaluasi multi-metrik, dapat ditarik beberapa kesimpulan sebagai berikut:

a. Implementasi *agent deep reinforcement learning* menggunakan algoritma *proximal policy optimization* (*PPO*) dengan teknik *reward shaping* pada *Environment Humanoid*-v5 berhasil dilakukan menggunakan *framework Stable-Baseline* stb3 berbasis *PyTorch*. *Reward shaping* dirancang dengan tiga komponen berbasis domain yaitu stabilitas postur (w1=1,0), efisiensi energi (w2=0,5), dan ketahanan gerak (w3=0,1) yang ditambahkan di atas *reward* bawaan *environment*.

b. PPO Standar sebagai baseline terbukti belum optimal dalam eksperimen ini, ditandai dengan mean reward yang hanya mencapai 1.215,92, episode length yang singkat sebesar 222,8 langkah, serta ketidakmampuan menghasilkan nilai *torso stability* yang terukur. Mengindikasikan tren ketidakstabilan yang konsisten selama proses pelatihan.

   Penerapan reward shaping terbukti secara signifikan meningkatkan performa agent PPO dibandingkan PPO Standar. Hal ini dibuktikan dengan peningkatan mean reward sebesar 611,0% dari 1.215,92 menjadi 8.641,51, serta peningkatan episode length sebesar 320,2% dari 222,8 menjadi 934,96

langkah. Selain itu, PPO dengan reward shaping mencapai performa terbaiknya lebih awal pada timestep ke-38.889.108 dibandingkan PPO Standar yang baru mencapainya pada timestep ke-44.200.000, menunjukkan bahwa reward shaping tidak hanya meningkatkan kualitas performa tetapi juga mempercepat proses konvergensi.Evaluasi multi-metrik yang digunakan, seluruhnya menunjukkan keunggulan *PPO* dengan *reward shaping*. Metrik torso *stability* hanya berhasil terukur pada skenario *reward shaping* (nilai 1,363) sementara *PPO* Standar tidak menghasilkan nilai yang terukur, membuktikan bahwa tanpa komponen stabilitas eksplisit *agent* tidak memiliki insentif untuk menjaga postur tubuh. Efisiensi energi *PPO* dengan *reward shaping* juga lebih baik dengan nilai 0,1316 dibandingkan *PPO* Standar sebesar 0,2436, menunjukkan pola gerak yang lebih halus dan hemat energi.

5.2 Saran

Berdasarkan hasil penelitian yang dilakukan masih memiliki sejumlah keterbatasan. Oleh karena itu terdapat beberapa saran sebagai bahan pertimbangan untuk penelitian berikutnya, yaitu:

a. Penelitian ini hanya menguji satu konfigurasi koefisien (w1=1,0; w2=0,5; w3=0,1). Untuk melakukan pencarian koefisien secara sistematis disarankan menggunakan metode seperti *grid search* atau *bayesian optimization* agar kombinasi bobot yang paling optimal dapat ditemukan.
b. Eksperimen *reward shaping* pada penelitian ini hanya dilakukan pada satu *seed* (*seed*=42). Untuk meningkatkan validitas dan reliabilitas hasil, dapat

menggunakan minimal 3-5 *seed* yang berbeda sehingga variansi hasil dapat dianalisis secara statistik.

c. Dapat menambahkan metrik seperti *forward velocity*, *fall count*, dan *smoothness of motion* untuk mendapatkan gambaran performa *agent* yang lebih lengkap dan realistis.

d. Penelitian ini terbatas pada *environment humanoid*-v5. Disarankan untuk menguji pendekatan *reward shaping* yang sama pada *environment* lain seperti *HalfCheetah*, *Ant*, atau Walker2D untuk mengevaluasi generalisasi metode yang dikembangkan.

e. Seluruh eksperimen dilakukan dalam lingkungan simulasi *MuJoCo*. Penelitian selanjutnya disarankan untuk mengeksplorasi teknik domain *randomization* atau transfer *Learning* agar model yang dihasilkan dapat diterapkan pada robot fisik nyata dengan lebih baik.

DAFTAR PUSTAKA

Akbar, R., Siroj, R. A., Win Afgani, M., & Islam Negeri Raden Fatah Palembang Abstract, U. (n.d.). Experimental Reseacrch Dalam Metodologi Pendidikan. Jurnal Ilmiah Wahana Pendidikan, Januari, 2023(2), 465–474. https://doi.org/10.5281/zenodo.7579001

Asadi, A., Chatterjee, K., Kafshdar Goharshady, E., Karrabi, M., & Shafiee, A. (2025). Qualitative Analysis of ω-Regular Objectives on Robust MDPs. arXiv preprint arXiv:2505.04539. https://doi.org/10.48550/arXiv.2505.04539

Ayabe, S., Otomo, T., Kera, H., & Kawamoto, K. (2025). Robustness evaluation of offline reinforcement learning for robot control against action perturbations. International Journal of Advanced Robotic Systems, 22(4). https://doi.org/10.1177/17298806251360454

Slamet, F. A. (2022). *Model Penelitian Pengembangan (R n D)*. Institut Agama Islam Sunan Kalijogo Malang.

Ghasemi, M., & Ebrahimi, D. (2024*)*. Introduction to Reinforcement Learning. arXiv preprint arXiv:2408.07712. https://doi.org/10.48550/arXiv.2408.07712

Ibrahim, S., Mostafa, M., Jnadi, A., Salloum, H., & Osinenko, P. (2024). Comprehensive Overview of Reward Engineering and Shaping in Advancing Reinforcement Learning Applications. *IEEE Access*, *12*, 175473–175500. https://doi.org/10.1109/ACCESS.2024.3504735

Jeon, S. H., Heim, S., Khazoom, C., & Kim, S. (2023). Benchmarking Potential Based Rewards for Learning Humanoid Locomotion. 2023 IEEE International Conference on Robotics and Automation (ICRA)*,* 9204–9210. https://doi.org/10.1109/ICRA48891.2023.10160885

Jeon, S. H., Heim, S., Khazoom, C., & Kim, S. (2023). Benchmarking Potential Based Rewards for Learning Humanoid Locomotion. *2023* IEEE International Conference on Robotics and Automation (ICRA), 9204–9210. https://doi.org/10.1109/ICRA48891.2023.10160885

Ji, X., & Li, G. (2023). Regret-Optimal Model-Free Reinforcement Learning for Discounted MDPs with Short Burn-In Time. Advances in Neural Information Processing Systems*, 36*, 80674–80689. https://doi.org/10.48550/arXiv.2305.15546

Kim, W., Shin, Y., Park, J., & Sung, Y. (2023). Sample-Efficient and Safe Deep Reinforcement Learning via Reset Deep Ensemble Agents. Advances in

Neural Information Processing Systems, *36*, 53239–53260. https://doi.org/10.48550/arXiv.2310.20287

Liu, S. D., Chen, C., & Zhang, S. (2025). Efficient Multi-Policy Evaluation for Reinforcement Learning. Proceedings of the AAAI Conference on Artificial Intelligence, *39*(18), 18951–18959. https://doi.org/10.1609/aaai.v39i18.34086

Murphy, K. (2024). *Reinforcement Learning: An Overview*. arXiv preprint arXiv:2412.05265. https://doi.org/10.48550/arXiv.2412.05265

Sasso, R., Conserva, M., & Rauber, P. (2023). Posterior Sampling for Deep Reinforcement Learning. Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 30042–30061. https://doi.org/10.48550/arXiv.2305.00477

Terven, J. (2025). Deep Reinforcement Learning: A Chronological Overview and Methods. *AI*, *6*(3), 46. https://doi.org/10.3390/ai6030046

V, J. K., & Elumalai, V. K. (2025). A proximal policy optimization based deep reinforcement learning framework for tracking control of a flexible robotic manipulator. Results in Engineering, 25, 104178. https://doi.org/10.1016/J.RINENG.2025.104178

Wang, S., Si, N., Blanchet, J., Zhou, Z., & Zhang, Z. (2024). Sample Complexity of Variance-Reduced Distributionally Robust Q-Learning. In Journal of Machine Learning Research (Vol. 25). http://jmlr.org/papers/v25/23-0526.html.

Zangirolami, V., & Borrotti, M. (2024). Dealing with uncertainty: balancing exploration and exploitation in deep recurrent reinforcement learning. arXiv preprint arXiv:2310.08331. https://doi.org/10.48550/arXiv.2310.08331

Zhang, L., Shen, L., Yang, L., Chen, S., Wang, X., Yuan, B., & Tao, D. (2022). Penalized Proximal Policy Optimization for Safe Reinforcement Learning. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence *(IJCAI-22)*, 3744–3750. https://doi.org/10.24963/ijcai.2022/520

Zhang, Z., Xu, J., Zhuang, Z., Zhang, H., Liu, J., Wang, D., & Zhang, S. (2024). A dynamical clipping approach with task feedback for Proximal Policy Optimization. arXiv preprint arXiv:2312.07624. https://doi.org/10.48550/arXiv.2312.07624

Zouitine, A., Bertoin, D., Clavier, P., Geist, M., & Rachelson, E. (2024). RRLS: Robust Reinforcement Learning Suite. arXiv preprint arXiv:2406.08406. https://doi.org/10.48550/arXiv.2406.08406

LAMPIRAN

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
Jl. G. Obos No.114 Telp.0536-3224593, 3225515 Fax.0536-3225515 Palangka Raya
email : humas@stmikplk.ac.id - website : www.stmikplk.ac.id

SURAT TUGAS PEMBIMBING TUGAS AKHIR

NO. 234/STMIK-3.C.2/AU/IX/2025

Ketua Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya menugaskan nama-nama tersebut di bawah ini :

1. Nama : Lili Rusdiana, M.Kom
   NIK : 198707282011007
   Sebagai Pembimbing I dalam **PEMBUATAN PROGRAM**
2. Nama : Catharina Elmayantie, M.Pd
   NIK : 197610252015003
   Sebagai Pembimbing II dalam **PENULISAN TUGAS AKHIR**

Untuk membimbing Tugas Akhir mahasiswa :

Nama	:	Rio Juliannur Saputra
NIM	:	C2255201075
Program Studi	:	TEKNIK INFORMATIKA (55201)
Tanggal Daftar	:	17/09/2025
Judul Tugas Akhir	:	Implementasi Agen Deep Reinforcement Learning Menggunakan Proximal Policy Optimization (PPO) dengan Reward Shaping dan Evaluasi Multi-Metrik pada Environment Humanoid

Demikian surat ini dibuat agar dapat dipergunakan sebagaimana mestinya, dan dilaksanakan dengan penuh tanggung jawab.

Palangka Raya, 26 September 2025

Ketua Program Studi,

26 September

Ir. ABDUL HADI, S.T., M.Kom
NIK. 198505102021101

TEMBUSAN :
1. Pembimbing I dan II
2. Mahasiswa yang bersangkutan
3. Arsip Prodi TI

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
Jl. G. Obos No.114 Telp.0536-3224593, 3225515 Fax.0536-3225515 Palangka Raya
email : humas@stmikplk.ac.id - website : www.stmikplk.ac.id

SURAT TUGAS PENGUJI TUGAS AKHIR

481/STMIK-3.C.2/AU/IV/2026

Ketua Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya menugaskan kepada nama- nama berikut :

1. Nama : Veny Cahya Hardita, M.Kom.
   NIK : 199504302020002
   Sebagai Ketua
2. Nama : Rudini, M.Pd
   NIK : 198709172015105
   Sebagai Sekretaris
3. Nama : Dr. Maura. Widyaningsih, S.Kom., M.Cs
   NIK : 197502232009002
   Sebagai Anggota
4. Nama : Lili Rusdiana, M.Kom
   NIK : 198707282011007
   Sebagai Anggota
5. Nama : Catharina Elmayantie, M.Pd
   NIK : 197610252015003
   Sebagai Anggota

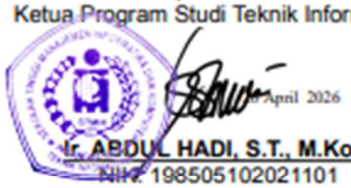
Untuk menguji Tugas Akhir mahasiswa :

Nama Mahasiswa	:	Rio Juliannur Saputra
N I M	:	C2255201075
Hari / Tanggal Ujian	:	Sabtu, 2 Mei 2026
Waktu	:	08.00 WIB
Judul TA	:	IMPLEMENTASI AGENT DEEP REINFORCEMENT LEARNING MENGGUNAKAN PROXIMAL POLICY OPTIMIZATION DENGAN REWARD SHAPING DAN EVALUASI MULTI-METRIK PADA ENVIRONMENT HUMANOID
Ruang	:	TI 3

Demikian surat ini dibuat agar dapat dipergunakan sebagaimana mestinya dan dilaksanakan dengan penuh tanggung jawab.

Palangka Raya, 20 April 2026

Ketua Program Studi Teknik Informatika,



Ir. ABDUL HADI, S.T., M.Kom
NIK. 198505102021101

Tembusan :
1. Dosen Penguji
2. Mahasiswa yang Bersangkutan
3. Arsip

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
Jl. G. Obos No.114 Telp.0536-3224593, 3225515 Fax.0536-3225515 Palangkaraya
email : humas@stmikplk.ac.id - website : www.stmikplk.ac.id

KARTU KEGIATAN KONSULTASI TUGAS AKHIR

Nama Mahasiswa : Rio Dhanner Saputra
NIM : C2355201075
Tanggal Persetujuan Judul : 17 / 09 / 2025
Judul Tugas Akhir : Implementasi Agen Deep Reinforcement Learning Menggunakan Proximal policy Optimization dengan Reward Shaping dan Evaluasi Multi-Metrik pada Environment Humanoid.

No.	Tanggal Konsultasi		Uraian	Tanda Tangan
	Terima	Kembali		
	26/9 25	26/9 25	Bab 1 ditengkapi & Perbaiki. Ikuti format dari prodi	
	27/9 25	27/9 25	Bab 2 lengkapi dan lanjut Bab 3.	
	20/09 25	20/09 25	Diskusi Bab 3 (Jenis, desain, dan instrumen penelitian)	
	26/09 25	26/09 25	Penyesuaian Tahapan R&D dan jadwal Penelitian.	
	09/10 25	09/10 25	Review dan finalisasi Bab I - III (ACC Seminar).	
	30/9 25	30/9 25	Revisi awal naskah proposal sesuai catatan dan Pedoman Penulisan TA.	
	2/10 25	2/10 25	Perbaiki Format Bab 1 judul, Latar belakang dan rumusan masalah	
	3/10 25	3/10 25	Konsultasi kelengkapan lampiran dan perbaikan minor.	
	4/10 25	4/10 25	(ACC Seminar).	
	4/2 2026	9/2 2026	Sesuaikan topik & Laporan	
	25/2 2026	25/2 2026	Lanjutkan Program	
	10/3 2026	12/3 2026	Lanjutkan penulisan dan demonstrasi	
	2/4 2026	2/4 2026	ACC Demo	

No.	Tanggal Konsultasi		Uraian	Tanda Tangan
	Terima	Kembali		
	13/4 2026		Acc Laporan & Lanjutkan ke dabang 2	
	14/4		Acc Sidang	
	14/4 2026	15/4 2026	Perbaikan Inti sari dan Abstract.	
	15/4 2026	16/4 2026	Lengkapi naskah	
	16/4 2026	17/4 2026	Lengkapi lampiran	
	17/4 2026	18/4 2026	ACC Sidang	

Menyetujui :

Dosen Pembimbing I,

Liki Kurniawan

Dosen Pembimbing II,

Catharina Elmayantie, M.Pd.

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
Jl. G. Obos No.114 Telp.0536-3224593, 3225515 Fax.0536-3225515 Palangka Raya
email : humas@stmikplk.ac.id - website : www.stmikplk.ac.id

SURAT TUGAS PEMBIMBING TUGAS AKHIR

NO. 234/STMIK-3.C.2/AU/IX/2025

Ketua Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya menugaskan nama-nama tersebut di bawah ini :

1. Nama : Lili Rusdiana, M.Kom
   NIK : 198707282011007
   Sebagai Pembimbing I dalam **PEMBUATAN PROGRAM**
2. Nama : Catharina Elmayantie, M.Pd
   NIK : 197610252015003
   Sebagai Pembimbing II dalam **PENULISAN TUGAS AKHIR**

Untuk membimbing Tugas Akhir mahasiswa :

Nama	:	Rio Juliannur Saputra
NIM	:	C2255201075
Program Studi	:	TEKNIK INFORMATIKA (55201)
Tanggal Daftar	:	17/09/2025
Judul Tugas Akhir	:	Implementasi Agen Deep Reinforcement Learning Menggunakan Proximal Policy Optimization (PPO) dengan Reward Shaping dan Evaluasi Multi-Metrik pada Environment Humanoid

Demikian surat ini dibuat agar dapat dipergunakan sebagaimana mestinya, dan dilaksanakan dengan penuh tanggung jawab.

Palangka Raya, 26 September 2025

Ketua Program Studi,

26 September

Ir. ABDUL HADI, S.T., M.Kom
NIK. 198505102021101

TEMBUSAN :
1. Pembimbing I dan II
2. Mahasiswa yang bersangkutan
3. Arsip Prodi TI