Whitepaper



KISLEK

Künstliche Intelligenz zur Steuerung verbundener Verkehrsknoten







Gefördert durch:
Bundesministerium für Digitales und Verkehr
aufgrund eines Beschlusses des Deutschen Bundestages

Abstract

Urbane Verkehrssteuerung ist heute vor allem dort herausfordernd, wo sich mehrere Lichtsignalanlagen (LSA) entlang eines stark belasteten Straßenzugs gegenseitig beeinflussen. Fahrzeuge, die an einer Kreuzung freigegeben werden, treffen bereits kurze Zeit später an der nächsten Anlage ein. Rückstaus wirken auf vorgelagerte Knoten zurück, Grünzeiten konkurrieren über mehrere Richtungen hinweg, und lokale Verbesserungen können den Verkehrsfluss an anderer Stelle beeinträchtigen. In solchen Situationen reicht es nicht aus, eine einzelne Kreuzung isoliert zu optimieren. **Benötigt wird eine Steuerung, die den Verbund mehrerer Knoten als zusammenhängendes System versteht.**

Genau hier setzt das Projekt **KISLEK – Künstliche Intelligenz zur Steuerung verbundener Verkehrsknoten –** an. Ziel war es, den im Vorgängerprojekt KI4LSA[1,2] erfolgreich erprobten Einsatz von Reinforcement Learning (RL) für eine einzelne reale LSA auf einen **Verbund mehrerer verkehrlich gekoppelter Knoten** zu übertragen und eine holistische Steuerung des Verkehrsflusses zu realisieren. Untersucht wurden drei miteinander verbundene LSA entlang der **Columbusstraße / B212 in der Innenstadt von Bremerhaven**: LSA 223, 224 und 225. Für diesen Abschnitt stand ein bereits kalibriertes mikroskopisches Verkehrsmodell in **VISSIM** zur Verfügung, ergänzt um reale Verkehrsdaten und eine im Realbetrieb bewährte koordinierte Referenzsteuerung.

Im Zentrum stand die Entwicklung **einer Multi-Agenten-RL-Architektur**, bei der jede der drei LSA durch einen eigenen KI-Agenten gesteuert wird. Dabei wurde die bestehende verkehrstechnisch geprüfte Steuerlogik als **aktiver Safety Layer** integriert – dies garantiert die RiLSA-Konformität und die Einhaltung aller Sicherheitsvorgaben. Die KI kann nur solche Folgephasen anfordern, die im jeweiligen Zustand zulässig sind. Zwischenzeiten, Mindestfreigaben und regulatorische Vorgaben bleiben in der vorhandenen Steuerung verankert. **Damit verbindet KISLEK KI-basierte Optimierung mit einem klar definierten sicherheits- und regelkonformen Handlungsrahmen.**

Ein wesentlicher Beitrag des Projekts liegt in seiner **Praxisnähe**. Training und Evaluation erfolgten vollständig in **VISSIM**, dem in der Verkehrsplanung etablierten Industriestandard für mikroskopische Verkehrssimulation. Es wurde ein leistungsfähiges Framework zur Simulation und zum Training von Verbundsteuerungen entwickelt. Verwendet wurde dabei lediglich vorhandene oder leicht nachrüstbare Sensorik – **Induktionsschleifen und Stauzähler**. Bereits mit diesen einfachen Detektionsgrößen gelang das Erlernen der Verbundsteuerung unter **Gewährleistung RiLSA-konformer Abläufe**. Im Unterschied zu vielen Forschungsarbeiten adressiert KISLEK damit direkt die Frage, wie KI unter realitätsnahen Bedingungen in bestehende Planungs- und Betriebsprozesse integriert werden kann.

Die Ergebnisse zeigen, dass eine **sichere, lernfähige und koordinierte KI-Steuerung für verbundene LSA technisch realisierbar ist**. Im finalen Stand bleibt die RL-Steuerung netzweit hinter der bereits sehr leistungsfähigen Referenzsteuerung zurück (u. a. +22 % mittlere Fahrzeugverlustzeit), zeigt aber einen stetigen Lernfortschritt und erreicht an einzelnen Knoten und Fahrströmen Referenzniveau oder übertrifft es (z. B. LSA 225 mit −1,2 %). Mit zunehmender Trainingsdauer nähert sich die KI dem Niveau der optimierten Referenzsteuerung an. Bereits zwischen mittleren und späten Trainingsständen sank die netzweite mittlere Fahrzeugverlustzeit deutlich. Einzelne Knoten und Fahrströme erreichten Referenzniveau oder verbesserten dieses leicht. Weitere Verbesserungen können durch längere Trainingszeiten, eine gezielte Optimierung der Belohnungsfunktion und Hyperparameter sowie einen weiterentwickelten Informationsaustausch zwischen den Agenten erzielt werden.

Die Ergebnisse stellen einen **großen Schritt in Richtung industrieller Anwendung** dar. Wissenschaftlich wurde eine Referenzarchitektur für sicherheitskonformes Multi-Agent-RL

[1] https://www.iosb-ina.fraunhofer.de/de/geschaeftsbereiche/maschinelles-lernen/forschungsthemen-und-projekte/projekt-KI4LSA.html

[2] https://www.iosb-ina.fraunhofer.de/content/dam/iosb/iosb-ina/documents/WhitePaper_KI4LSA.pdf

geschaffen. Das entwickelte Framework bildet die Grundlage für kommunale Pilotprojekte, Produktentwicklung und Feldtests. Perspektivisch kann die Technologie weniger Stau, kürzere Reisezeiten und weniger Emissionen bei hoher Anpassungsfähigkeit an wechselnde Verkehrsmuster ermöglichen.

KISLEK liefert damit keinen vereinfachten Labornachweis, sondern einen belastbaren Beitrag zur Frage, wie KI für die optimierte Verkehrssteuerung **sicher, anschlussfähig und industrienah** entwickelt werden kann. Das Projekt zeigt, dass der Erfolg solcher Systeme nicht allein vom Lernalgorithmus abhängt. Ebenso entscheidend sind Sicherheitslogik, Schnittstellen, Simulationsumgebung, Datenbasis, Bewertungsmethodik und die Fähigkeit, sich in bestehende Verkehrssteuerungsarchitekturen einzufügen.

1. Der stark frequentierte Verkehrsverbund, die verkehrstechnisch optimierte Referenzsteuerung und der KISLEK-Ansatz

KISLEK steht für *Künstliche Intelligenz zur Steuerung verbundener Verkehrsknoten*. Im Projekt wurde untersucht, wie ein KI-basiertes Steuerungsverfahren für mehrere verkehrlich gekoppelte LSA entwickelt, abgesichert und unter realitätsnahen Bedingungen bewertet werden kann. Konkret betrachtete das Projekt drei Lichtsignalanlagen (LSA 223, 224 und 225) entlang einer neuralgischen Verkehrszelle der Columbusstraße / B212 in der Innenstadt von Bremerhaven (siehe Abbildung 1).

Diese drei Knoten bilden einen relevanten Teil eines stark belasteten innerstädtischen Streckenzugs, auf dem eine koordinierte Signalsteuerung für die Aufrechterhaltung des Verkehrsflusses wesentlich ist. Ohne funktionierende Koordinierung würde der Verkehr in der Spitzenstunde in diesem Bereich regelmäßig an seine Grenzen stoßen.



Abbildung 1: Draufsicht des im Projekt betrachteten Knotenverbunds in Bremerhaven. Kartengrundlage: © OpenStreetMap-Mitwirkende, ODbL.:

Die klassische Verkehrssteuerung berücksichtigt Wechselwirkungen über Knoten im Verbund bereits über Koordinierung, Signalprogramme und grüne Wellen. Bei stark schwankender Nachfrage, mehreren gleichzeitig zu berücksichtigenden Zielgrößen und dynamischen

Rückstauwirkungen stoßen diese Verfahren jedoch an Grenzen. Adaptive und datengetriebene Ansätze können hier eine sinnvolle Ergänzung sein.

Eine besondere Stärke des Projekts lag in seiner Ausgangsbasis. Für den untersuchten Abschnitt lag bereits ein kalibriertes mikroskopisches VISSIM-Modell vor, das von der BERNARD Gruppe im Rahmen vorangegangener verkehrsplanerischer Arbeiten aufgebaut wurde und insbesondere auch Fußgänger- und Radverkehrsströmen berücksichtigt. Zusätzlich standen reale Verkehrszähldaten für die Nachmittagsspitze sowie die bestehende verkehrstechnisch optimierte Referenzsteuerung zur Verfügung. Das Projekt musste also nicht bei einem abstrakten Demonstrator beginnen, sondern konnte auf einer fachlich belastbaren und praxisnahen Grundlage aufsetzen.

Ebenso wichtig ist die Qualität der Vergleichsbasis. Die entwickelte RL-Steuerung trat nicht gegen eine einfache Festzeitsteuerung oder eine bewusst schwache Baseline an, sondern gegen eine bereits leistungsfähige, koordinierte Referenzsteuerung, die im Realbetrieb erfolgreich eingesetzt wird. Gerade dadurch haben die Ergebnisse ein hohes Gewicht: Wenn sich eine KI unter diesen Bedingungen dem Referenzniveau annähert oder dieses in Teilbereichen erreicht, ist das wesentlich aussagekräftiger als ein Vergleich mit einer wenig optimierten Ausgangslösung.

Das Projekt wurde als Verbund von Fraunhofer IOSB-INA und BERNARD Gruppe ZT GmbH umgesetzt. Die assoziierten Partner PTV Group, BASt, Straßen.NRW sowie die Seestadt Bremerhaven als Betreiberin der Anlagen unterstützten das Vorhaben begleitend mit fachlichem Austausch. Während Fraunhofer insbesondere für Konzeption, RL-Architektur, Training und wissenschaftliche Auswertung verantwortlich war, brachte BERNARD das Verkehrsmodell, die verkehrsplanerische Bewertung und die bestehende Steuerungskompetenz in das Projekt ein. Unterstützt wurde die Arbeit zusätzlich durch fachlichen Austausch mit weiteren Akteuren aus Simulation, Verwaltung und Infrastrukturumfeld.

KISLEK verfolgte dabei im Kern fünf Leitfragen (siehe Tabelle 1):

Tabelle 1: Leitfragen des Projekts KISLEK.

Leitfrage	**Inhalt**
Algorithmische Architektur	*Ist ein Single-Agent- oder Multi-Agent-Ansatz geeigneter?*
Informationsaustausch	*Welche Nachbarinformationen braucht koordinierte Steuerung?*
Reward Engineering	*Wie wird lokales Lernen mit netzweitem Nutzen verbunden?*
Sicherheit und Regelkonformität	*Wie kann KI innerhalb eines zulässigen Steuerungsrahmens agieren?*
Einfache Detektionsgrößen	*Wie kann die KI mit Detektionsgrößen, die in der Regel an LSA vorhanden (Induktionsschleifen) oder leicht nachrüstbar (Stauzähler) sind, eine Verbundsteuerung erlernen?*

In der wissenschaftlichen Literatur zu RL für Lichtsignalsteuerung gibt es in den letzten Jahren eine sehr hohe Dynamik. Viele Arbeiten zeigen beeindruckende Verbesserungen in Simulationsumgebungen und belegen, dass lernbasierte Verfahren grundsätzlich in der Lage sind, Verkehrssteuerungsprobleme zu adressieren. Gleichzeitig bleibt bei vielen dieser Ansätze offen, wie weit ihre Ergebnisse tatsächlich in reale Verkehrssteuerung übertragbar sind. Häufig basieren sie auf vereinfachten Netzmodellen, abstrahierten Steuergeräten, reduzierten Verkehrsarten oder einem Aktionsraum, der sicherheitsrelevante Randbedingungen nicht realistisch abbildet.

KISLEK setzt genau an dieser Lücke an. Das Projekt ist nicht primär darauf ausgelegt, unter idealisierten Bedingungen möglichst hohe relative Verbesserungswerte zu erzielen. Ziel ist vielmehr, eine **anschlussfähige, sichere und belastbar evaluierte Lösung** zu entwickeln, die reale Anforderungen erfüllt.

Der erste wesentliche Unterschied liegt in der **Simulationsumgebung**. Training und Evaluation erfolgten vollständig in **VISSIM**. VISSIM ist in Verkehrsplanung und Verkehrstechnik seit Jahren etabliert und wird von Planungsbüros, Kommunen und Straßenbaulastträgern als Industriestandard eingesetzt. Das ist aus zwei Gründen relevant: Erstens erhöht es die Realitätsnähe der Untersuchung. Zweitens verbessert es die Übertragbarkeit der Ergebnisse in reale Planungs- und Genehmigungsprozesse.

Der zweite zentrale Unterschied liegt in der **Sicherheitsarchitektur**. In KISLEK wurde die bestehende verkehrstechnisch geprüfte Steuerung nicht ersetzt, sondern bewusst als Bestandteil des Systems erhalten. Die KI darf also nicht beliebig schalten, sondern nur solche Folgephasen anfordern, die auf Basis der vorhandenen Steuerlogik im jeweiligen Zustand zulässig sind. Dadurch bleiben Zwischenzeiten, Mindestfreigabezeiten und andere sicherheitsrelevante Anforderungen fest im System verankert. Diese Form der Integration schafft eine Brücke zwischen lernender Optimierung und regelgebundener Verkehrssteuerung.

Zusammengefasst verschiebt KISLEK den Fokus von der Frage „Kann KI prinzipiell steuern?“ hin zu der deutlich praxisnäheren Frage:

Wie muss KI entwickelt und eingebettet werden, damit sie in realen Verkehrssteuerungssystemen überhaupt sinnvoll einsetzbar wird?

2. Die KI-basierte Verbundsteuerung

2.1 Design

Reinforcement Learning (RL) bietet hierfür einen adaptiven Ansatz: Die Agenten lernen durch wiederholte Interaktion mit der Simulationsumgebung, welche zulässigen Phasenanforderungen in bestimmten Verkehrssituationen günstig sind. Dies ist insbesondere bei verkehrlich gekoppelten LSA relevant, weil sich die Wirkungen eines Signalwechsels zeitlich verzögert in Rückstaus, Reisezeiten und der Belastung benachbarter Knoten zeigen. Die Entscheidungen werden auf Grundlage aktueller Beobachtungen des Verkehrsgeschehens getroffen.

Zu Beginn des Projekts wurde zunächst ein **Single-Agent-Ansatz** betrachtet. Die Idee dahinter war, den gesamten Verbund aus drei LSA durch einen zentralen Agenten steuern zu lassen. Im Projektverlauf zeigte sich jedoch, dass ein einzelner Agent die unterschiedlichen Knotentopologien, Nachfrageverteilungen und lokalen Besonderheiten nicht robust genug abbilden konnte. Die Evaluation sowie die Analyse der Lernmetriken deuteten darauf hin, dass ein zentraler Agent Schwierigkeiten hatte, die notwendige Spezialisierung für alle drei Knoten gleichzeitig zu erreichen. Herausforderungen sind dabei insbesondere die unterschiedlichen Knotentopologien und Verkehrsdynamiken.

Deshalb wurde auf einen **Multi-Agenten-Ansatz** umgestellt. In dieser Architektur wird jede LSA durch einen eigenen Agenten gesteuert. Jeder Agent besitzt ein eigenes neuronales Netz und lernt eine knotenspezifische Strategie. Gleichzeitig wurden Beobachtungs- und Aktionsräume so strukturiert, dass die Grundarchitektur über alle drei Agenten hinweg konsistent bleibt. Das erleichtert die technische Implementierung, erlaubt aber dennoch individuelle Lernstrategien pro Knoten.

Dieser Schritt war für das Projekt zentral. Denn er macht deutlich, dass sich die Steuerung verbundener LSA nicht einfach als vergrößerte Variante der Einzelknotensteuerung verstehen lässt. Vielmehr braucht sie eine Architektur, die sowohl lokale Spezialisierung als auch koordiniertes Verhalten ermöglicht.

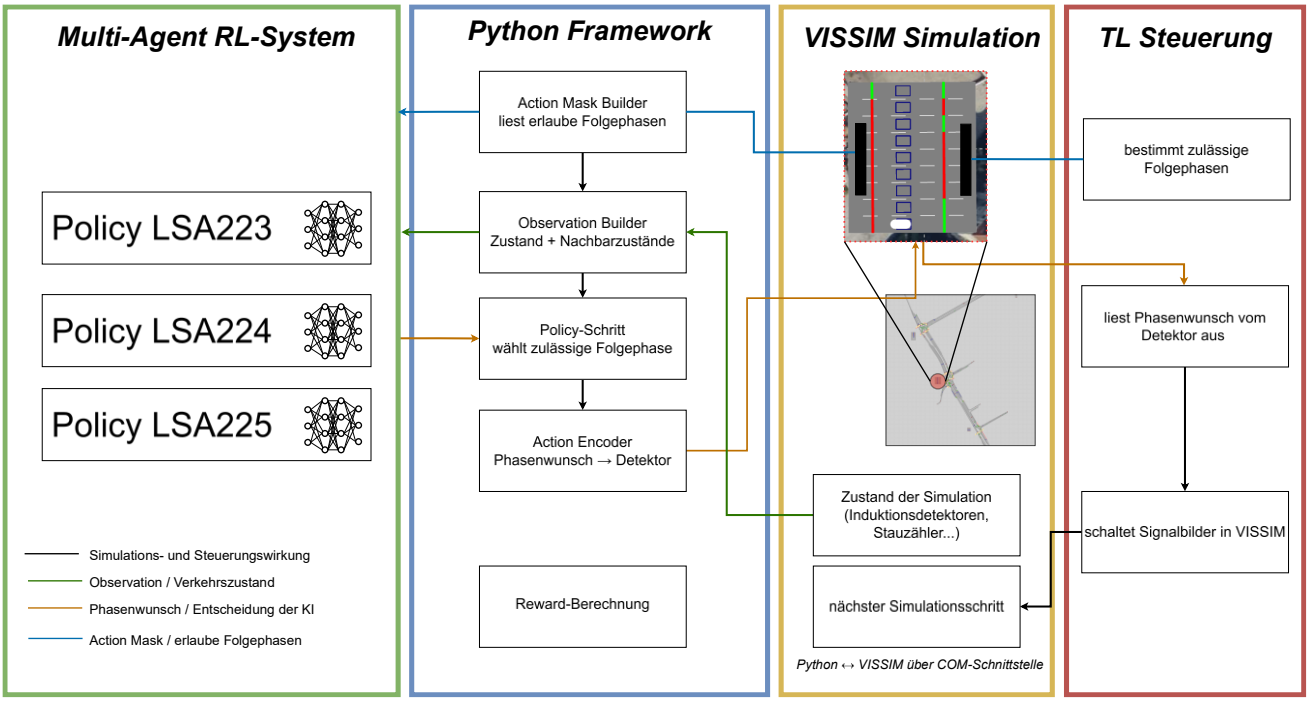


Abbildung 2: KISLEK-Architektur: Multi-Agent-RL mit VISSIM-COM und TL-Referenzsteuerung

Jeder Agent erhält pro Sekunde einen strukturierten Beobachtungsvektor, der den aktuellen Zustand seiner LSA und des unmittelbaren Umfelds beschreibt. Dazu gehören insbesondere:

- Belegungen von Induktionsdetektoren,
- normierte Staulängen auf den relevanten Fahrstreifen,
- die aktuell aktive Phase,
- die Zeit seit dem letzten Phasenwechsel,
- Informationen wie lange belastete Phasen nicht bedient wurden,
- Informationen über benachbarte Knoten, insbesondere deren Phasen und Staubelastung.

Wichtig ist, dass sich diese Informationsauswahl bewusst an Daten orientiert, die auch im realen Einsatz bereits verfügbar oder mit vertretbarem Aufwand nachrüstbar wären. Neben den bestehenden Induktionsschleifen wurden in der Simulation zusätzliche Stauzähler verwendet, die sich auch in realen Anwendungen implementieren lassen. Damit wurde vermieden, dass die KI auf idealisierte oder in der Realität praktisch nicht verfügbare Zustandsgrößen angewiesen ist.

Ein zentraler Punkt der Systemgestaltung war die bewusste Begrenzung des Handlungsraums. Die KI schaltet **nicht direkt einzelne Signalbilder oder Signalgruppen frei**. Stattdessen fordert sie in jedem Zeitschritt eine gewünschte **Folgephase** an. Welche Folgephasen in einem bestimmten Moment zulässig sind, wird durch die bestehende Steuerungslogik vorgegeben.

Diese Entscheidung ist aus zwei Gründen wichtig. Erstens bleibt der Handlungsraum verkehrstechnisch strukturiert und beherrschbar. Zweitens wird vermieden, dass die KI

sicherheitskritische Zustandswechsel selbst erlernen und verhindern muss. Das Lernen konzentriert sich damit auf die strategische Auswahl sinnvoller Phasen im Rahmen einer sicheren Steuerungslogik.

Die Koordination zwischen den Agenten erfolgt in KISLEK nicht durch explizite direkte Kommunikation im Sinne eines Nachrichtenaustauschs zwischen den Netzwerken. Stattdessen wird koordiniertes Verhalten über zwei Mechanismen gefördert:

1. **Nachbarinformationen im Beobachtungsraum** Jeder Agent erhält Informationen über die relevanten benachbarten Knoten. So „sieht“ LSA 223 den Zustand von LSA 224, LSA 224 sieht sowohl LSA 223 als auch LSA 225, und LSA 225 sieht wiederum LSA 224.
2. **Ein globaler Anteil in der Reward-Funktion** Zusätzlich zum lokalen Ziel enthält die Belohnungsfunktion einen globalen Anteil, der den Zustand des gesamten Verbunds berücksichtigt. Dadurch wird verhindert, dass Agenten ausschließlich lokale Gewinne maximieren und dabei Probleme an benachbarte Knoten verlagern.

Gerade diese Kombination aus lokalem Beobachtungsraum und globalem Lernsignal ist für die Verbundsteuerung entscheidend. Sie erlaubt den Agenten, aus ihrer lokalen Perspektive zu handeln, ohne das Gesamtsystem vollständig aus dem Blick zu verlieren.

Konkret erhält jeder Agent pro Zeitschritt einen strukturierten Beobachtungsvektor (siehe Tabelle 2), der die folgenden Informationen enthält:

Tabelle 2: Beobachtungsvektor des Reinforcement Learning.

Block	Information	Wert
Loop-Detektor-Features	Belegungsflag je Induktionsschleife	Float [0–1], 0 = keine Belegung, 1 = 100 % Belegung im Messintervall
Stauzähler-Features	Staulänge normiert auf Referenzlänge	Float [0–1], 0 = kein Stau, 1 = Stau über gesamte Referenzlänge
Detektor-Masken	Gibt an, welche Detektor-Slots gültig belegt sind, One-Hot	Binär (0/1), 1 = reale Detektorposition, 0 = Platzhalter / nicht belegt
Phase	Aktuell aktive Phase, One-Hot	Binär (0/1), One-Hot; 1 = aktuell aktive Phase, 0 = inaktiv
Congested Steps per Phase	Normierte Anzahl aufeinanderfolgender Schritte mit Stau je inaktive Phase	Float [0–1], 0 = kein anhaltender Stau, 1 ≈ ≥300 s Stau bei roter Phase
Zeit seit letztem Phasenwechsel	Normiert auf 180 s	Float [0–1], 0 = gerade gewechselt, 1 ≈ ≥180 s seit letztem Phasenwechsel
Phasenstatus	known-Flag und Transition-Flag, One-Hot	Binär (0/1), 2 Flags: known=1 → Phase eindeutig bekannt; transition=1 → Phasenübergang
Nachbar-Features	Staulänge	Float [0–1], 0 = kein Stau am Nachbarknoten, 1 = starke Staubelastung am Nachbarknoten

	Phase One-Hot der Nachbarkreuzungen (max. 2)	Binär (0/1), One-Hot; 1 = aktuell aktive Phase des Nachbarknotens
Agent	Identifikation des Agenten (LSA 223/224/225), One-Hot	Binär (0/1), One-Hot über {223,224,225}; 1 an Position der betrachteten LSA, sonst 0

2.2 Sicherheitsarchitektur

Eine der wichtigsten Projektentscheidungen in KISLEK war, die KI **nicht als vollständigen Ersatz** der bestehenden verkehrstechnischen Steuerung zu verstehen. In sicherheitskritischen Infrastrukturen wie LSA ist diese Unterscheidung grundlegend. Es reicht nicht aus, dass ein lernendes System im Durchschnitt gute Ergebnisse erzielt. Es muss jederzeit sichergestellt sein, dass alle sicherheits- und regelrelevanten Randbedingungen eingehalten werden.

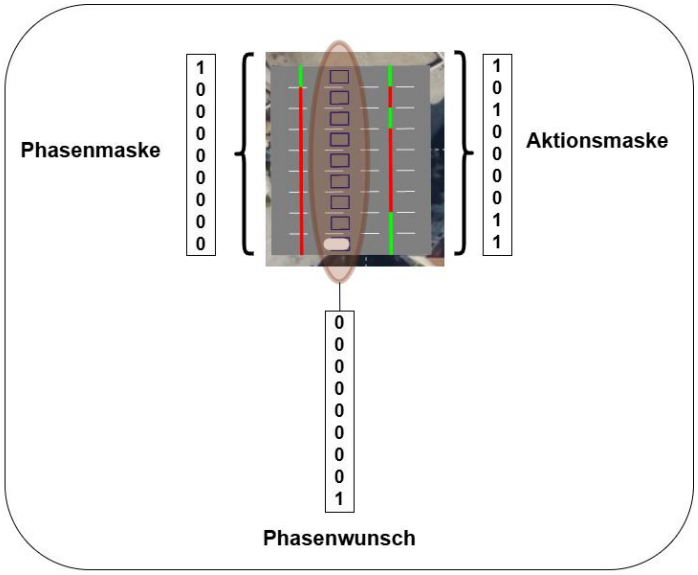


Abbildung 3: Methodik der Kommunikation VISSIM zu KI-Modul

Deshalb wurde die bestehende Steuerlogik als **aktive Sicherheitsschicht** in die Architektur integriert. Diese Logik bleibt zuständig für:

- zulässige Phasenfolgen,
- Zwischenzeiten,
- Mindestfreigabezeiten,
- Mindestsperrzeiten,
- weitere regulatorisch relevante Randbedingungen.

Die KI darf also nur innerhalb eines bereits verifizierten und verkehrstechnisch zulässigen Regelrahmens handeln. Technisch wurde dies über eine **Action-Masking-Logik** umgesetzt. Zu jedem Entscheidungszeitpunkt wird aus der bestehenden Steuerung abgeleitet, welche Folgephasen aktuell zulässig sind. Nur diese Optionen werden dem Agenten als erlaubte Aktionen zur Verfügung gestellt. Unzulässige Aktionen sind algorithmisch ausgeschlossen.

Diese Architektur ist mehr als ein technisches Detail. Sie ist eine zentrale Voraussetzung für spätere Praxistauglichkeit. In vielen wissenschaftlichen Demonstratoren wird Sicherheit

implizit vereinfacht oder außerhalb des eigentlichen Lernproblems behandelt. KISLEK zeigt dagegen, dass es möglich ist, die KI von Anfang an so einzubetten, dass sie mit bestehenden Sicherheitsmechanismen zusammenarbeitet.

Da keine direkte variablenbasierte bidirektionale Schnittstelle zur Realisierung der Kommunikation zwischen KI-Modul und Steuerung in der benötigten Form verfügbar war, wurde ein robuster Kommunikationsweg über VISSIM-Objekte entwickelt (siehe Abbildung 3). Die KI meldet Phasenwünsche über definierte Kommunikationsdetektoren an die Steuerung. Die Steuerung meldet im Gegenzug über zusätzliche Signalzustände zurück, welche Zielphasen aktuell zulässig sind. Damit entsteht ein geschlossener Kommunikationskreis, in dem die KI nicht außerhalb der Steuerungslogik operiert, sondern mit ihr gekoppelt ist.

Für die Praxis ist genau das die zentrale Botschaft:

Eine KI-basierte LSA-Steuerung muss nicht als unkontrollierte Black Box gedacht werden, sondern kann innerhalb eines abgesicherten und überprüfbaren technischen Rahmens Verkehrsflüsse optimieren.

2.3 Training und Simulationsumgebung

Ursprünglich war im Projekt ein Dual-Simulator-Ansatz vorgesehen. SUMO sollte als effiziente Trainingsumgebung dienen, während VISSIM für die abschließende Evaluation genutzt werden sollte. Diese Aufteilung erschien zunächst sinnvoll, weil SUMO in der RL-Forschung weit verbreitet ist und paralleles Training oft leichter ermöglicht. Im Projektverlauf zeigte sich jedoch, dass dieser Ansatz für KISLEK mit erheblichen Nachteilen verbunden gewesen wäre.

Erstens ließ sich das von BERNARD bereits kalibrierte VISSIM-Modell nicht ohne Weiteres verlustfrei nach SUMO übertragen. Dadurch wäre ein **Sim-to-Sim-Gap** entstanden, also eine Differenz zwischen Trainings- und Evaluationswelt. Zweitens ließ sich die sicherheitsrelevante Steuerungslogik über die Yunex-API zum Projektzeitpunkt nicht mit mehreren parallelen Simulationsinstanzen betreiben, sodass der zentrale Vorteil von SUMO, das parallelisierbare Training, entfallen wäre. Drittens hätte die Synchronisierung zwischen externer Steuerung, KI-Modul und Simulationsumgebung zusätzlichen technischen Aufwand und Fehlerquellen erzeugt.

Vor diesem Hintergrund wurde entschieden, Training und Evaluation **vollständig in VISSIM** durchzuführen. Diese Entscheidung war technisch anspruchsvoller, aber inhaltlich konsequent. Sie vermeidet Transferrisiken zwischen unterschiedlichen Simulationsumgebungen und sorgt dafür, dass Training und Bewertung auf derselben realitätsnahen Basis erfolgen. Allerdings hatte dieser Weg auch Konsequenzen: Das Training wurde deutlich zeitintensiver als in stärker parallelisierbaren Umgebungen.

Das Training basierte auf dem PPO-Algorithmus aus Ray RLlib. Jede Episode umfasste einen simulierten Zeitraum von 2.000 Sekunden, ergänzt um ein Warm-up, damit sich zunächst ein realistischer Verkehrszustand aufbauen konnte. Entscheidungen wurden im Sekundentakt getroffen. Während des Trainings wurden regelmäßig Checkpoints gespeichert, um Lernfortschritte und Zwischenstände systematisch auswerten zu können.

Gerade für Verkehrsplanungsbüros und Infrastrukturverantwortliche ist dieser Punkt von Bedeutung: KISLEK zeigt, dass RL **nicht nur in akademischen Standardumgebungen**, sondern auch in einer industrienahen Simulationsarchitektur betrieben werden kann. Das erhöht die Anschlussfähigkeit an reale Planungs- und Betriebsprozesse erheblich.

2.4 Trainingskonfiguration

Das Training der Agenten erfolgte mit dem PPO-Algorithmus aus Ray RLlib auf Basis der VISSIM-Trainingsumgebung. Aufgrund der technischen Randbedingungen der VISSIM-COM-Schnittstelle, konnte das Training nicht effizient parallelisiert werden, sondern musste sequentiell mit einer einzelnen VISSIM-Instanz durchgeführt werden. Für jede der drei LSA wurde im

Multi-Policy-Ansatz eine eigene Policy trainiert. Die drei neuronalen Netze besitzen dieselbe Architektur, weil Beobachtungs- und Aktionsvektor bei allen Agenten dieselbe Länge haben. Die Semantik der Beobachtungs- und Aktionsräume ist LSA-spezifisch, so dass die Agenten knotenspezifische Steuerungsstrategien lernen können; Struktur und Implementierung sind für alle drei LSA jedoch einheitlich.

Jede Trainingsepisode basierte auf einem simulierten Zeitabschnitt von 2.000 Sekunden. Vor Beginn der eigentlichen Bewertung innerhalb einer Episode wurde ein Warm-up von 300 s durchgeführt, damit sich zunächst ein realistischer Verkehrszustand im Netz einstellen konnte. Die Entscheidungsfrequenz der Agenten betrug 1 s, in der für jede LSA eine neue Beobachtung erzeugt wird.

Das verwendete neuronale Netz bestand aus einem mehrschichtigen vollständig verbundenen Netz, wobei nach jeder linearen Schicht Layer Normalization, eine ReLU-Aktivierung und Dropout erfolgen. Auf dem Merkmalsrepräsentationsblock bauen ein Policy-Head zur Berechnung der Aktionslogits und ein Value-Head zur Schätzung des Zustandswerts auf. Die Action Mask wurde direkt auf die Aktionslogits angewendet, sodass unzulässige Phasenfolgen während Training und Inferenz nicht ausgewählt werden konnten.

Für das PPO-Training wurden pro Optimierungsschritt Trainingsbatches mit 10.000 Zeitschritten verwendet, die in Minibatches der Größe 1.024 unterteilt wurden. Für jedes Trainingsbatch wurden zehn PPO-Optimierungsepochen durchgeführt. Als zentrale Hyperparameter wurden ein Diskontierungsfaktor von 0,97, eine Lernrate von $1 \cdot 10^{-4}$, Generalized Advantage Estimation mit 0,95, ein PPO-Clipping-Parameter von 0,2, ein Entropiekoeffizient von 0,005, ein Value-Loss-Gewicht von 0,5 sowie ein Gradient-Clipping von 5,0 verwendet. Diese Parameter stellten sich im Projektverlauf als geeignet heraus und ermöglichen eine stabile Lernentwicklung und ausreichend exploratives Verhalten bei gleichzeitig regelkonformer Phasenwahl.

Zur Nachvollziehbarkeit und zur späteren Auswahl geeigneter Modelle wurden während des Trainings regelmäßig Checkpoints gespeichert. In der finalen Konfiguration erfolgte die Speicherung alle zwei Trainingsiterationen. Zusätzlich wurde die Option genutzt, Trainingsläufe aus bestehenden Checkpoints fortzusetzen, was den Transfer von gelernten Zwischenständen ermöglicht. Für die abschließende Evaluation wurden die austrainierten Agenten (Lernkurve konvergiert) in der VISSIM-Umgebung gegen die konventionelle Referenzsteuerung verglichen.

2.5 Reward-Logik und verkehrliche Zielsetzung

RL funktioniert nur dann sinnvoll, wenn das Lernziel angemessen beschrieben wird. In KISLEK war genau dieses sogenannte **Reward Engineering** ein zentrales Entwicklungsfeld. Die Herausforderung bestand darin, ein Lernsignal zu definieren, das nicht nur lokale Entlastung an einer einzelnen Anlage belohnt, sondern eine robuste und faire Steuerung des gesamten Verbunds fördert.

Die finale Reward-Funktion kombiniert daher mehrere Zielkomponenten:

- Reduktion des Rückstaus,
- Verringerung hoher Detektorbelegungen,
- Vermeidung dauerhaft nicht bedienter belasteter Phasen,
- Förderung einer ausgewogeneren Belastungsverteilung über Fahrstreifen,
- Berücksichtigung des Gesamtzustands des Verbunds.

Der wichtigste Einzelterm ist die **Staulänge**, weil sie direkt mit verkehrlicher Überlastung zusammenhängt. Ergänzt wird sie durch Detektorbelegung, um auch kurzfristige Belastungen und Annäherungen sichtbar zu machen. Ein zusätzlicher Phasen-Strafterm verhindert, dass Agenten einzelne Phasen mit vorhandener Nachfrage dauerhaft vernachlässigen. Gerade in Nebenrichtungen oder bei Fußgängerquerungen wäre ein solches Verhalten verkehrlich nicht

akzeptabel. Schließlich adressiert ein Fairness-Term die Ungleichverteilung von Rückstau über mehrere Zufahrten hinweg.

Entscheidend ist aber die Kombination des **lokalen** und **globalen** Reward (vgl. Abbildung 4). Jeder Agent soll in erster Linie seine eigene Anlage gut steuern. Gleichzeitig darf er lokale Verbesserungen nicht auf Kosten benachbarter Knoten erzielen. Deshalb enthält die Reward-Funktion einen bewusst gesetzten globalen Anteil, der den Zustand des gesamten Verbunds widerspiegelt. Dieses Design soll kooperatives Verhalten fördern, ohne die lokale Verantwortlichkeit der einzelnen Agenten aufzulösen.

Aus verkehrsplanerischer Sicht ist dabei wichtig, dass die Reward-Funktion nicht direkt die HBS-Bewertung (Handbuch für die Bemessung von Straßenverkehrsanlagen) oder die mittlere Fahrzeugverlustzeit optimiert. Diese Größen liegen erst nach einer vollständigen Simulation aggregiert vor und eignen sich daher nicht als unmittelbares Schritt-für-Schritt-Lernsignal. Stattdessen folgt KISLEK dem Prinzip des **Reward Shaping**: Es werden Größen optimiert, die während des laufenden Betriebs messbar sind und mit hoher Wahrscheinlichkeit zu besseren verkehrlichen Zielgrößen führen. Genau diese Verbindung aus technischer Lernlogik und verkehrlicher Interpretierbarkeit war ein wesentlicher Bestandteil des Projekterfolgs.

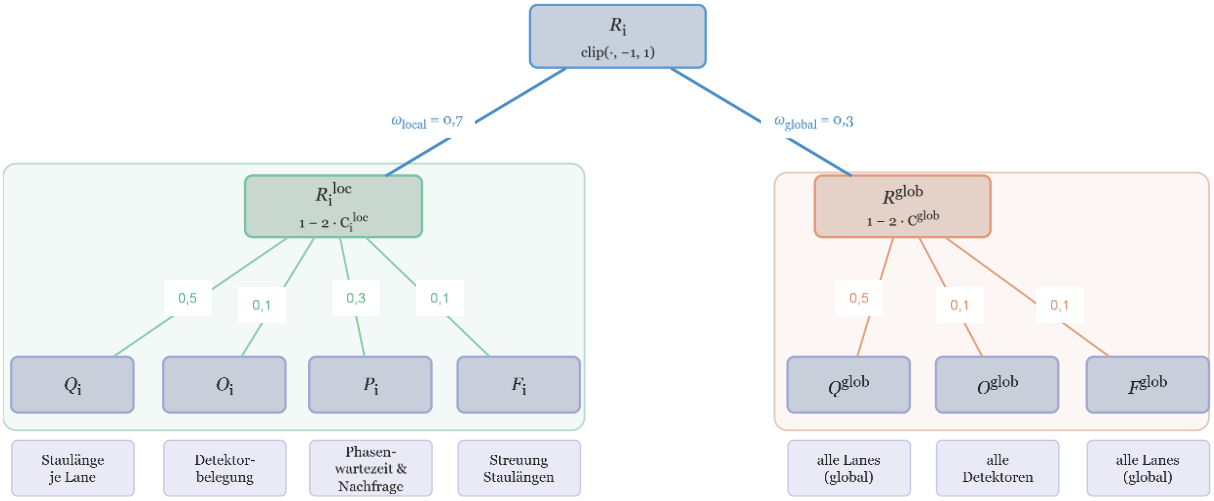


Abbildung 4: Formelbaum der Reward-Funktion mit Gewichten.

2.6 Formalisierung der Reward Funktion

Die finale Reward-Funktion kombiniert lokale und globale Kostenanteile. Für jede RL-gesteuerte LSA $i \in \mathcal{I}$ werden zunächst lokale Kosten C_{loc} berechnet. Diese setzen sich aus vier normierten Teilgrößen zusammen:

$$C_i^{\mathrm{loc}} = \omega_q Q_i + \omega_o O_i + \omega_p P_i + \omega_f F_i.$$

Dabei beschreibt Q_i die normierte Staulänge am Knoten, O_i die mittlere Belegung der Induktionsdetektoren, P_i einen Phasen-Strafterm für länger nicht bediente, aber belastete Phasen und F_i einen Fairness-Term zur Bewertung der Ungleichverteilung der Staulängen über die betrachteten Fahrstreifen. Aus den lokalen Kosten wird ein lokaler Reward im Wertebereich nahe $[-1{,}1]$ gebildet:

$$R_i^{\mathrm{loc}} = 1 - 2C_i^{\mathrm{loc}}.$$

Ein Wert nahe 1 entspricht damit einem günstigen Verkehrszustand mit geringen Kosten, während niedrigere Werte nahe -1 eine zunehmende verkehrliche Belastung anzeigen.

Zusätzlich wird ein globaler Kostenanteil über alle RL-gesteuerten LSA berechnet. Dieser globale Term berücksichtigt die netzweite Staulänge, die mittlere Belegung und die globale Ungleichverteilung der Staulängen:

$$C^{\text{glob}} = \omega_q Q^{\text{glob}} + \omega_o O^{\text{glob}} + \omega_f F^{\text{glob}},$$
$$R^{\text{glob}} = 1 - 2C^{\text{glob}}.$$

Der globale Term dient dazu, kooperatives Verhalten zwischen den Agenten zu fördern. Ohne diesen Anteil besteht bei Multi-Agent-Systemen die Gefahr, dass einzelne Agenten lokal günstige Entscheidungen treffen, die Rückstaus oder Verzögerungen an benachbarte Knoten verlagern. Durch die Berücksichtigung globaler Kosten erhält jeder Agent zusätzlich eine Rückmeldung über den Zustand des gesamten betrachteten LSA-Verbunds.

Der finale Reward für Knoten i ergibt sich als gewichtete Kombination aus lokalem und globalem Reward:

$$R_i = \text{clip}\left(\omega_{\text{local}} R_i^{\text{loc}} + \omega_{\text{global}} R^{\text{glob}}, -1{,}1\right).$$

Die Funktion $\text{clip}(x, -1{,}1)$ begrenzt den resultierenden Wert auf das Intervall $[-1{,}1]$, um numerisch stabile Lernsignale zu erhalten. Der globale Kostenanteil nutzt die Gewichte für Staulänge, Belegung und Fairness (Summe 0,8), sodass der globale Reward im Bereich [−0,6, 1] liegt. In der finalen Implementierung wurden folgende Gewichte verwendet:

$$\omega_q = 0.5, \quad \omega_o = 0.1, \quad \omega_p = 0.3, \omega_f = 0.1,$$
$$\omega_{\text{local}} = 0.7, \quad \omega_{\text{global}} = 0.3.$$

Damit liegt der Schwerpunkt der Optimierung auf der Reduzierung von Rückstau. Gleichzeitig werden Detektorbelegung, Phasenbalance und Fairness als zusätzliche Stabilisierungsterme berücksichtigt. Der lokale Reward dominiert bewusst gegenüber dem globalen Anteil, da jeder Agent primär für die konkrete Steuerungsqualität seiner eigenen LSA verantwortlich ist. Der globale Reward ergänzt dieses Signal um eine kooperative Komponente und reduziert die Wahrscheinlichkeit, dass lokale Verbesserungen zu Verschlechterungen im Verbund führen.

Der Queue-Term Q_i basiert auf den Stauzählern der betrachteten Fahrstreifen. Für jede Lane ℓ eines Knotens i wird die aktuelle Staulänge $s_{i,\ell}$ auf eine Referenz- bzw. Maximallänge $m_{i,\ell}$ normiert:

$$g_{i,\ell} = \text{clip}\left(\frac{s_{i,\ell}}{m_{i,\ell}}, 0{,}1\right).$$

Der lokale Queue-Term ergibt sich anschließend als Mittelwert über alle betrachteten Lanes des Knotens:

$$Q_i = \frac{1}{|\mathcal{L}i|} \sum_{\ell \in \mathcal{L}i} g_{i,\ell}$$

Der globale Queue-Term Q^{glob} wird analog über alle betrachteten Lanes aller RL-gesteuerten Knoten gebildet:

$$Q^{\mathrm{glob}} = \frac{1}{|\mathcal{L}^{\mathrm{glob}}|} \sum_{(i,\ell)\in\mathcal{L}^{\mathrm{glob}}} g_{i,\ell}$$

Dieser Term stellt sicher, dass die tatsächliche Rückstausituation im Netz direkt in das Lernsignal eingeht. Die lane-basierte Normierung verhindert dabei, dass einzelne lange oder kurze Zufahrten die Kosten unverhältnismäßig stark dominieren.

Der Occupancy-Term O_i bewertet die mittlere Belegung der Induktionsdetektoren am jeweiligen Knoten. Für eine Menge von Detektoren $d \in \mathcal{D}_i$ und eine prozentuale Belegung $\mathrm{Occ}_{i,d}$ ergibt sich:

$$O_i = \frac{1}{|\mathcal{D}_i|} \sum_{d\in\mathcal{D}_i} \frac{\mathrm{Occ}_{i,d}}{100}$$

Der globale Occupancy-Term wird als Mittelwert über alle lokalen Occupancy-Werte berechnet:

$$O^{\mathrm{glob}} = \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} O_i.$$

Die Detektorbelegung ergänzt die Stauzähler, da sie auch kurzfristige Belastungen und Annäherungen erfassen kann. Während die Staulänge insbesondere Rückstau abbildet, liefert die Belegung zusätzliche Information über die aktuelle Auslastung der Detektionsbereiche.

Der Fairness-Term F_i wurde eingeführt, um stark ungleich verteilte Rückstaus innerhalb eines Knotens zu bestrafen. Dies ist insbesondere relevant, wenn einzelne Fahrstreifen über längere Zeit stark belastet sind, während andere Zufahrten regelmäßig bedient werden, obwohl sie weniger belastet sind. Für einen Knoten mit $n_i = |\mathcal{L}_i|$ betrachteten Lanes wird zunächst der Mittelwert der normierten Staulängen berechnet:

$$\bar{g}_i = \frac{1}{n_i} \sum_{\ell\in\mathcal{L}_i} g_{i,\ell}$$

Die lokale Fairness-Strafe ist dann proportional zur Streuung der normierten Staulängen:

$$F_i = \mathrm{clip}\left(\frac{2}{\sqrt{n_i}} \sqrt{\sum_{\ell\in\mathcal{L}_i} \left(g_{i,\ell} - \bar{g}_i\right)^2}\,, 0{,}1 \right)$$

Für Knoten mit weniger als zwei betrachteten Lanes wird $F_i = 0$ gesetzt. Ein niedriger Wert bedeutet eine gleichmäßigere Auslastung, ein hoher Wert zeigt eine starke Ungleichverteilung der Rückstaus an. Der globale Fairness-Term F^{glob} wird analog über alle Lanes im gesamten betrachteten Verbund gebildet.

Der Phasen-Strafterm P_i adressiert eine spezifische Herausforderung RL-basierter LSA-Steuerung: Ohne zusätzliche Gegenmaßnahme kann ein Agent lernen, einzelne Phasen sehr lange nicht zu bedienen, wenn dies kurzfristig den lokalen Reward verbessert. Dies kann

insbesondere bei Nebenrichtungen, Fußgängerquerungen oder schwächer belasteten Zufahrten zu nicht akzeptablem Verhalten führen. Daher wird für jede Phase $\varphi \in \Phi_i$ betrachtet, ob dieser Phase eine relevante Staulänge zugeordnet ist und wie lange diese Belastung bereits konsekutiv besteht.

Sei $q_{i,\varphi}$ die einer Phase zugeordnete Staulänge in Metern (Summe über die zugehörigen Stauzähler) und $k_{i,\varphi}$ die Anzahl aufeinanderfolgender Sekunden, in denen die Phase nicht aktiv war und die zugehörige Staulänge 3 m überstieg. Zunächst wird ein Indikator definiert, ob eine relevante Queue vorhanden ist (Staulänge über 1 m):

$$\mathbf{1}_{i,\varphi}^{\text{queue}} = \begin{cases} 1, q_{i,\varphi} > 1 \\ 0, \text{sonst.} \end{cases}$$

Die zeitliche Persistenz der Staubelastung wird über eine stückweise Funktion $f(k)$ bewertet. In der finalen Konfiguration wurde hierfür $K_{\text{SAT}} = 90s$ verwendet: Die Strafe steigt damit in den ersten 90 s linear an, erreicht dort 0,5 und springt anschließend auf den vollen Wert 1.

$$f(k) = \begin{cases} \dfrac{0.5\,k}{K_{\text{SAT}}}, & 0 \le k < K_{\text{SAT}} \\ 1, & k \ge K_{\text{SAT}} \end{cases}$$

Der Beitrag einer Phase lautet damit:

$$p_{i,\varphi} = \mathbf{1}_{i,\varphi}^{\text{queue}} \cdot f\left(k_{i,\varphi}\right)$$

Der Phasen-Strafterm des Knotens ergibt sich als Mittelwert über alle Phasen:

$$P_i = \text{clip}\left(\frac{1}{|\Phi_i|}\sum_{\varphi \in \Phi_i} p_{i,\varphi}\,, 0{,}1\right)$$

Dieser Term wirkt als Phase-Starvation-Penalty. Er bestraft nicht das reine Nichtbedienen einer Phase, sondern nur das Nichtbedienen bei gleichzeitig vorhandener Nachfrage. Dadurch wird vermieden, dass leere Zufahrten unnötig priorisiert werden, während belastete, aber länger nicht bediente Phasen zunehmend stärker in das Lernsignal eingehen.

3. Evaluation

Die Evaluation der entwickelten RL-Steuerung erfolgte in der VISSIM-Umgebung auf Basis des kalibrierten Simulationsmodells der drei Knoten LSA 223, LSA 224 und LSA 225. Ziel der Evaluation war der direkte Vergleich der RL-gesteuerten Lösung mit BERNARDs konventioneller Referenzsteuerung. Diese Referenzsteuerung entspricht dem verkehrstechnischen, BERNARD-optimierten Ausgangszustand des betrachteten Streckenabschnitts und dient daher als zentrale Vergleichsbasis für alle Ergebnisgrößen. Beide Steuerungsvarianten wurden auf derselben Netzgeometrie, denselben Verkehrsnachfragewahrscheinlichkeitsdichten zur Nachmittagsspitze und denselben Messstellen ausgewertet, sodass die gemittelten Unterschiede in den Ergebnissen unmittelbar auf die jeweilige Steuerungsstrategie zurückgeführt werden können.

Zu berücksichtigen ist, dass die RL-Steuerung nicht gegen eine einfache oder historisch gewachsene Festzeitsteuerung antritt, sondern gegen eine bereits als „Top-Steuerung“ eingestufte, koordinierte Referenzsteuerung auf der B212. Nach Einschätzung der BERNARD Gruppe liegt diese im betrachteten Abschnitt nahe am verkehrstechnischen Optimum; die Evaluation ist daher eine harte Probe für die KI und zeigt, ob sie sich in einem anspruchsvollen, stark verknüpften Streckenzug behaupten kann.

Neben den austrainierten RL-Agenten, die im finalen RL-Checkpoint enthalten sind, wurden zusätzlich relevante Zwischenstände des Trainings evaluiert. Dadurch kann nicht nur die Leistungsfähigkeit der finalen RL-Steuerung bewertet werden, sondern auch nachvollzogen werden, ob und wie stark sich die Steuerungsqualität der RL-Agenten im Verlauf des Trainings verbessert.

Die Methodik orientiert sich an dem für die Verkehrsplanung etablierten HBS. Neben netzweiten Kennzahlen wurden auch knoten-, messstellen- und fahrstromspezifische Auswertungen durchgeführt. Dies ist insbesondere für einen LSA-Verbund relevant, da eine lokal verbesserte Steuerung an einem Knoten nicht übermäßig zu Lasten benachbarter Knoten gehen darf. Die Evaluation kombiniert daher klassische VISSIM-Kennzahlen, eine HBS-/QSV-Bewertung, Signalzeitenanalysen sowie RL-spezifische Diagnosegrößen.

Evaluationsprotokoll

Die Evaluation wurde für die Referenzsteuerung und die ausgewählten RL-Trainingsstände mit identischem Simulations- und Auswertungsprotokoll durchgeführt. Jeder Evaluationslauf bestand aus einer Vorlaufzeit zur Stabilisierung des Verkehrszustands und einem anschließenden Auswertezeitraum von einer Stunde Nachmittagsspitze. Während der Evaluation trifft jeder RL-Agent im Sekundentakt eine Entscheidung über die anzufordernde Folgephase seiner LSA. Das vollständige Evaluationsprotokoll ist in Tabelle 3 zusammengefasst.

Tabelle 3: Evaluationsprotokoll

Parameter	Wert / Beschreibung
Simulationsumgebung	VISSIM, kalibriertes Modell der Columbusstraße / B212
Steuerungsvarianten	Referenzsteuerung und RL-Steuerung
Betrachtete Knoten	LSA 223, LSA 224, LSA 225
Auswertezeitraum	3.600 s Simulationszeit
Vorlaufzeit (Warm-up)	900s
Entscheidungsintervall RL	1 s
Finaler Bewertungsstand	Trainingsiteration 108, in der sich eine Sättigung der Lernkurve stabilisierte
Anzahl Evaluationsläufe je Steuerungsvariante	10 Läufe mit unterschiedlichen Zufallsseeds (42 – 51)
Seed-Auswertung	Aggregation über die 10 Evaluationsläufe; ergänzend Best-Seed-Auswertung
Best-Seed-Kriterium	Minimaler mittlerer VEHDELAY
Vergleichsbasis	Konventionelle BERNARD-optimierte Referenzsteuerung

Zur Berücksichtigung der stochastischen Einflüsse der mikroskopischen Verkehrssimulation wurden für jede Steuerung jeweils 10 Simulationsläufe durchgeführt. Jeder Simulationslauf basiert auf einem Zufallsseed, so dass sich die Wahrscheinlichkeitsdichten der

Verkehrsanforderungen leicht, wie in der Realität, von Nachmittagsspitze zu Nachmittagsspitze unterscheiden. Die verschiedenen Steuerungen wurden jeweils mit den 10 selben Seeds evaluiert.

Die Ergebnisse wurden zunächst seedübergreifend aggregiert, indem Mittelwerte und Streuungen der zentralen Metriken ausgewertet wurden. Zusätzlich wurde für jeden betrachteten Trainingsstand eine Best-Seed-Auswertung durchgeführt. Als Auswahlkriterium wurde die mittlere Fahrzeugverlustzeit (VEHDELAY) verwendet. Diese Best-Seed-Auswertung dient als ergänzende Darstellung des leistungsfähigsten Simulationslaufs eines Steuerungsansatzes; die Bewertung der Robustheit erfolgt über die seedübergreifende Auswertung.

Evaluierte Trainingsstände und gesehene Simulationszeit

Neben den austrainierten RL-Agenten wurden ausgewählte Zwischenstände des Trainings separat evaluiert. Ziel dieser Zwischencheckpoint-Auswertung ist es, den Lernverlauf nicht ausschließlich anhand interner RL-Metriken wie Reward, Policy Loss oder Value Loss zu beurteilen, sondern auch anhand verkehrstechnischer Kennzahlen zu prüfen, wie sich die Steuerungsqualität im Verlauf des Trainings verbessert.

Die ausgewählten Checkpoints decken frühe, mittlere und späte Trainingsphasen ab. Sie wurden mit demselben Evaluationsprotokoll wie der finale Agent bewertet. Dadurch kann untersucht werden, wie die zunehmende Trainingsdauer zu einer systematischen Verbesserung der VISSIM-Kennzahlen führt oder ob einzelne Trainingsstände instabile bzw. verkehrlich ungünstige Strategien erzeugen.

Tabelle 4: Evaluierte Trainingsstände im Lernverlauf.

Bewerteter Stand	Simulationssekunden	Agenten-Entscheidungsschritte	Einordnung
RL-Checkpoint 2: Früher Zwischenstand	20.000	60.000	frühe Trainingsphase / Startverhalten
RL-Checkpoint 46: Mittlerer Zwischenstand	460.000	1.380.000	erste stabilisierte Lernphase
RL-Checkpoint 62: Mittlerer bis später Zwischenstand	620.000	1.860.000	fortgesetztes Training
RL-Checkpoint 82: Später Zwischenstand	810.000	2.430.000	fortgeschrittene Steuerungsstrategie
RL-Checkpoint 108: Finaler evaluierter Stand	1.030.000	3.090.000	abschließende Bewertung

Da drei RL-Agenten (für jede LSA ein Agent) gleichzeitig trainiert wurden, entspricht die Zahl der Agenten-Entscheidungsschritte dem Dreifachen der Umgebungsschritte. Bei einem Entscheidungsintervall von 1 s können die Umgebungsschritte als gesehene Simulationssekunden interpretiert werden. Für die Bewertung des Trainingsverlaufs wird daher sowohl die verkehrliche Entwicklung über die evaluierten Checkpoints als auch die interne Lernentwicklung des PPO-Trainings betrachtet.

Ausgewertete Metriken

Die verkehrliche Hauptbewertung erfolgt gemäß den Qualitätsstufen des Verkehrsablaufs nach HBS (FGSV 2015) und basiert auf den VISSIM-Metriken VEHDELAY, TRAVTM und QLEN. Der VEHDELAY ist dabei die zentrale Kenngröße, da er die mittlere Verzögerung der Fahrzeuge gegenüber einer ungestörten Fahrt beschreibt und direkt mit der verkehrlichen Qualität der Steuerung zusammenhängt. Die Reisezeit TRAVTM ergänzt diese Betrachtung auf Routenebene, während die Staulängen QLEN und QLENMAX die lokale Rückstausituation an den einzelnen Zufahrten und Stauzählern abbilden (siehe Tabelle 5).

Zusätzlich wurden die Qualitätsstufen des Verkehrsablaufs nach HBS bestimmt. Die HBS-/QSV-Auswertung wurde sowohl aggregiert auf Ebene der LSA als auch fahrstrombezogen für die relevanten Ströme der B212 durchgeführt. Dadurch kann bewertet werden, inwiefern Verbesserungen nicht nur netzweit auftreten, sondern auch auf den maßgebenden Hauptströmen B212 wirksam werden.

Ergänzend wurden die Grün- und Rotzeitverteilungen aus den LZV-Ausgaben von VISSIM analysiert. Diese Auswertung dient der Plausibilitätsprüfung der gelernten Steuerungsstrategie. Sie zeigt, wie die RL-Steuerung verkehrstechnisch nachvollziehbare Signalzeitmuster erzeugt oder wie einzelne Signalgruppen unplausibel bevorzugt bzw. benachteiligt werden.

Neben den klassischen VISSIM-Kennzahlen wurden RL-spezifische Diagnosegrößen ausgewertet. Dazu gehören der Gesamt-Reward, lokale und globale Reward-Anteile, die normierte Staulänge, Phasen-Strafterme, Fairness-Terme sowie agentenspezifische Phasenkennzahlen. Diese Größen sind nicht als primäre verkehrsplanerische Bewertungsmetriken zu verstehen, sondern dienen der Interpretation des Lernprozesses und der gelernten Steuerungsstrategie.

Dabei ist grundsätzlich festzuhalten, dass das eigentliche Bewertungsziel nach HBS, die Einordnung des Verkehrsablaufs in Qualitätsstufen auf Basis der mittleren Fahrzeugverzögerung, vom RL-Agenten nicht direkt erlernt werden kann. Die HBS-Qualitätsstufen stellen eine nachgelagerte, aggregierte Auswertung dar, die erst nach Abschluss einer Simulation berechnet werden kann und damit nicht als schrittzeitliches Rückkopplungssignal innerhalb des Lernprozesses verfügbar ist. Die eingesetzte Reward-Funktion verwendet daher Ersatzgrößen, die im laufenden Simulationsbetrieb messbar sind und einen kausalen Zusammenhang mit den HBS-Zielgrößen aufweisen: Normierte Staulängen, Fahrzeugbelegungen sowie Fairness- und Phasen-Strafterme sollen dabei die Entstehungsbedingungen hoher Fahrverzögerungen widerspiegeln. Der Reward-Ansatz folgt damit dem Prinzip des Reward Shaping: Anstelle des nicht direkt zugänglichen Ziels werden diejenigen Zwischengrößen belohnt, deren Optimierung das HBS-Ziel mit hoher Wahrscheinlichkeit begünstigt.

Tabelle 5: Ausgewertete Metriken

Metrik	**Beschreibung**	**Werkzeug / Datenquelle**
VEHDELAY	Mittlere Fahrzeugverlustzeit [s/Fzg] je Messstelle, je LSA und netzweit	VISSIM Verlustzeitmessung
TRAVELTIME	Mittlere Fahrzeugreisezeit [s]	VISSIM Reisezeitmessung
QLEN	Mittlere Staulänge [m] je Stauzähler	VISSIM Stauzähler
QLENMAX	Maximale Staulänge [m] je Stauzähler	VISSIM Stauzähler
QSTOPS	Anzahl bzw. Häufigkeit von Stauereignissen	VISSIM Stauzähler
QSV / LOS	Qualitätsstufe des Verkehrsablaufs nach HBS	VISSIM/HBS-basierte Auswertung

Grün-/Rotzeiten	Verteilung der Signalzeiten je Signalgruppe	VISSIM-Auswertung
Reward-Komponenten	Lokaler Reward, globaler Reward, Queue-Norm, Phase-Penalty, Fairness-Penalty	RL-Evaluationslogs
Phasenkennzahlen	Phasenwechsel, Phasenzeitanteile, maximale Phasenhaltezeiten	RL-Evaluationszusammenfassung

Vergleichsebenen und Ergebnisdarstellung

Für die Ergebnisdarstellung wurden zwei Vergleichsebenen unterschieden. Die erste Ebene bildet den verkehrlichen Hauptvergleich zwischen der konventionellen BERNARD-optimierten Referenzsteuerung und den austrainierten RL-Agenten. Diese Auswertung ist maßgebend für die abschließende Bewertung der entwickelten Steuerung.

Die zweite Ebene umfasst den Vergleich ausgewählter RL-Trainingsstände. Durch die Analyse der Trainingsstände kann aufgezeigt werden, wie sich die verkehrlichen Kennzahlen mit zunehmender Trainingsdauer systematisch verbessern, und es zeigt sich, dass einzelne Trainingsstände zwar hohe interne Rewards erzielen, aber keine entsprechende Verbesserung der VISSIM-basierten Verkehrskennzahlen zeigen.

Die Kennzahlen werden sowohl verbundweit als auch auf Ebene der einzelnen LSA und Messstellen dargestellt. Die netzweite Auswertung beantwortet die Frage, inwiefern der LSA-Verbund insgesamt profitiert. Die LSA-spezifische Auswertung zeigt, ob Verbesserungen gleichmäßig über die drei Knoten verteilt sind oder eher einzelne Knoten profitieren, während andere verschlechtert werden. Die messstellenspezifische Auswertung erlaubt schließlich eine räumlich detaillierte Analyse der betroffenen Zufahrten.

Zentrale Ergebnisübersicht

Die zentrale Ergebnisübersicht stellt die BERNARD-optimierte Referenzsteuerung den austrainierten RL-Agenten gegenüber. Sie fasst die wichtigsten netzweiten, LSA-spezifischen und HBS-orientierten Kennzahlen zusammen. Die Tabelle dient als Ausgangspunkt für die anschließende Detailanalyse.

Tabelle 6: Zentrale Ergebnisübersicht BERNARD-optimierte Referenzsteuerung vs. austrainierte RL-Agenten (RL-Steuerung).

Metrik	Referenzsteuerung	RL-Steuerung
Netzwerk VEHDELAY [s/Fzg]	30,94	37,77
Netzwerk TRAVTM [s]	49,41	56,26
Netzwerk QLEN [m]	8,46	9,60
VEHDELAY LSA 223 [s/Fzg]	15,62	17,63
VEHDELAY LSA 224 [s/Fzg]	36,71	46,67
VEHDELAY LSA 225 [s/Fzg]	26,57	26,25
Schlechteste QSV eines B212-Fahrstroms	E	F
Anzahl B212-Fahrströme mit QSV D–F	1	2
Mittlere Wartezeit B212-Fahrströme [s/Fzg]	22,95	29,57

In der Zusammenfassung (siehe Tabelle 6) zeigt die RL-Steuerung gegenüber der konventionellen BERNARD-optimierten Referenzsteuerung im netzweiten Mittel eine höhere mittlere Fahrzeugverlustzeit (+22 %), eine längere Reisezeit (+13,9 %) sowie größere Staulängen (+13,4 %). Die stärkste Steigerung der Verlustzeiten tritt an dem topologisch und

verkehrstechnisch anspruchsvollsten Knoten des Verbunds, der LSA 224 auf (+27,1 %), während an LSA 225 eine geringfügige Verbesserung erkennbar ist (−1,2 %). Die fahrstrombezogene B212-Auswertung zeigt eine Zunahme der mittleren Wartezeit sowie eine Verschlechterung der schlechtesten Qualitätsstufe von E auf F.

Entscheidend für die Einordnung dieser Ergebnisse ist der Blick auf den Lernverlauf (Tabelle 7): Die Zwischencheckpoint-Auswertung belegt eine systematische, deutliche Verbesserung der verkehrlichen Qualität im Trainingsverlauf. Der netzweite VEHDELAY sinkt von 58,1 s/Fzg bei Iteration 46 auf 37,8 s/Fzg beim austrainierten Agenten, eine Reduktion um über 35 % allein durch Trainingsdauer. Dieser Trend lässt auf weiteres Verbesserungspotenzial bei verlängertem Training schließen. Die RL-Agenten haben damit nachweislich grundlegende Strukturen der Verbundsteuerung erlernt und entwickeln sich klar in Richtung des Referenzniveaus (30,9 s/Fzg).

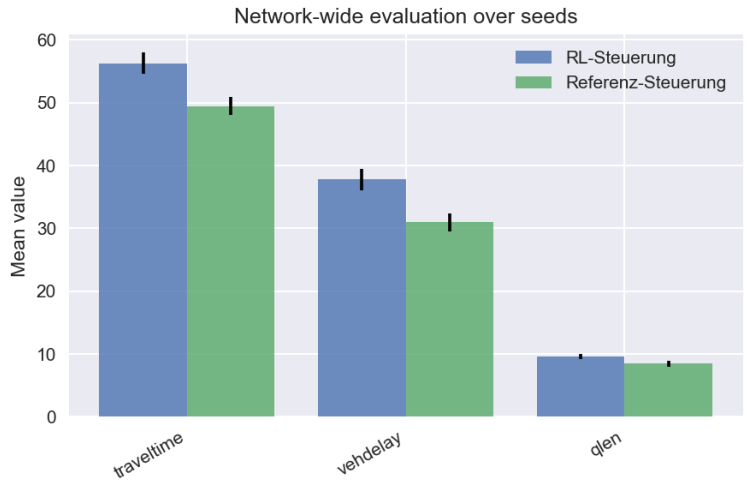


Abbildung 5: Vergleich der netzweiten VISSIM-Kennzahlen zwischen der Referenzsteuerung und den austrainierten RL-Agenten (RL-Steuerung). Dargestellt sind die zentralen Kenngrößen TRAVTM, VEHDELAY und QLEN.

Verkehrlicher Lernverlauf über die evaluierten Trainingsstände

Zur Bewertung des Lernverlaufs wurden die VISSIM-Kennzahlen der evaluierten Zwischenstände miteinander verglichen. Diese Auswertung zeigt, inwiefern sich die verkehrliche Wirkung der RL-Steuerung mit zunehmender Trainingsdauer verbessert. Dabei ist zu berücksichtigen, dass PPO als stochastisches Optimierungsverfahren keinen strikt monotonen Verbesserungsverlauf erwarten lässt. Einzelne Zwischenstände können daher schlechter abschneiden als frühere Checkpoints. Entscheidend ist, dass über den Trainingsverlauf eine plausible Tendenz zu einer besseren verkehrlichen Qualität erkennbar ist.

Abbildung 6 zeigt die Entwicklung der netzweiten VISSIM-Kennzahlen über die evaluierten Trainingsstände. Alle drei Metriken, VEHDELAY, TRAVTM und QLEN, weisen eine klare Verbesserungstendenz von frühen zu späten Trainingsständen auf. Der netzweite VEHDELAY sinkt von 58,1 s/Fzg (Iteration 46) über 49,9 s/Fzg (Iteration 62) und 39,8 s/Fzg (Iteration 82) auf 37,8 s/Fzg beim austrainierten Agenten. Trotz dieser deutlichen Verbesserung im Trai-

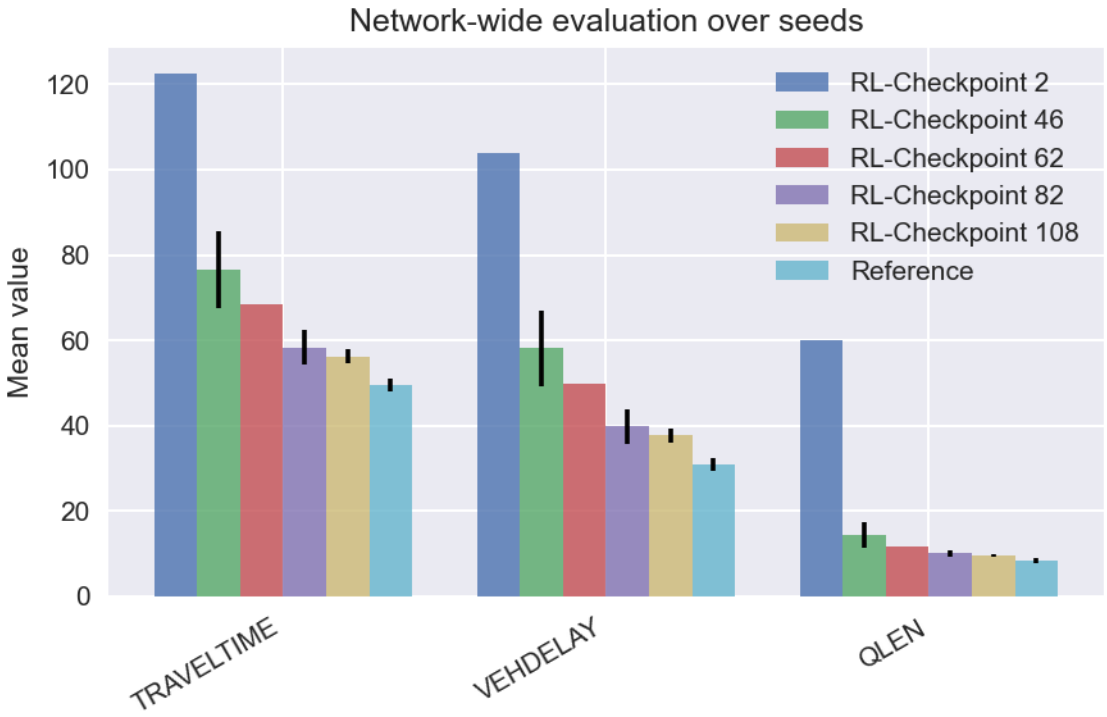


Abbildung 6: Entwicklung der netzweiten VISSIM-Kennzahlen über die evaluierten Trainingsstände. Dar-gestellt sind die seedübergreifenden Mittelwerte und Streuungen für die Referenzsteuerung sowie die RL-Checkpoints. Die Abbildung dient der Bewertung, ob sich die verkehrliche Qualität der RL-Steuerung im Trainingsverlauf verbessert.

ningsverlauf wird das Niveau der Referenzsteuerung (30,9 s/Fzg) noch nicht erreicht. Die Streuungen über die Seeds nehmen mit fortschreitendem Training tendenziell ab, was auf eine zunehmende Stabilisierung der gelernten Strategie hindeutet. Der nicht strikt monotone Verlauf zwischen einzelnen Checkpoints ist für das stochastische PPO-Verfahren, wie beschrieben, erwartbar und schränkt die Gesamtaussage nicht ein.

Neben der netzweiten Betrachtung ist die LSA-spezifische Auswertung erforderlich, da Verbesserungen im Verbund nicht gleichmäßig verteilt sein müssen. Ein netzweiter Vorteil kann beispielsweise dadurch entstehen, dass ein Knoten deutlich verbessert wird, während ein anderer Knoten unverändert bleibt oder sich gar verschlechtert.

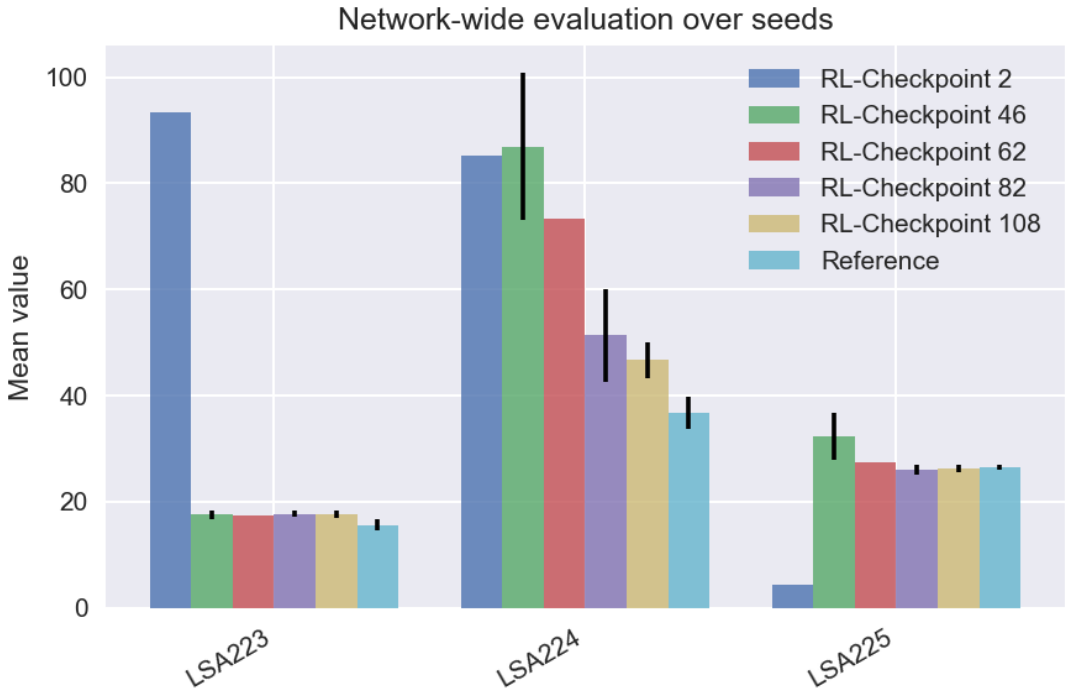


Abbildung 7: Entwicklung der mittleren Fahrzeugverlustzeit je LSA über die evaluierten Trainingsstände. Die Darstellung zeigt, ob Verbesserungen gleichmäßig über LSA 223, LSA 224 und LSA 225 auftreten oder ob einzelne Knoten besonders stark profitieren bzw. belastet werden.

Abbildung 7 verdeutlicht, dass der netzweite Lernfortschritt maßgeblich durch die Verbesserung an LSA 224 getrieben wird. Dort sinkt der VEHDELAY von 87,0 s/Fzg (Iteration 46) auf 46,7 s/Fzg beim austrainierten Agenten, ein Rückgang um mehr als 46 %. Dennoch verbleibt LSA 224 auch im finalen Stand mit 46,7 s/Fzg über dem Referenzwert von 36,7 s/Fzg, was die im Verbund topologisch und verkehrstechnisch anspruchsvollste Konstellation dieses Knotens zu bestätigen scheint. LSA 223 und LSA 225 stabilisieren sich ab dem mittleren Trainingsstand auf einem Niveau nahe der Referenz (LSA 225 mit 26,3 s/Fzg sogar leicht darunter, LSA 223 mit 17,6 s/Fzg rund 13 % darüber). Die LSA-spezifische Betrachtung bestätigt damit, dass die RL-Agenten knotenspezifisch unterschiedliche Lernfortschritte erzielen und die netzweite Verbesserung nicht auf eine gleichmäßige Wirkung über alle Knoten zurückzuführen ist.

Tabelle 7: Zusammenfassung des verkehrlichen Lernverlaufs

Trainingsstand	Netzwerk VEHDELAY [s/Fzg]	Netzwerk TRAVTM [s]	Netzwerk QLEN [m]	VEHDELAY LSA 223 [s/Fzg]	VEHDELAY LSA 224 [s/Fzg]	VEHDELAY LSA 225 [s/Fzg]	Einordnung
Referenz	30,94	49,41	8,46	15,62	36,71	26,57	Vergleichsbasis
RL-Checkpoint 2	103,87	122,40	60,17	93,25	85,12	4,26	Früher Zwischenstand

RL-Checkpoint 46	58,13	76,6	14,34	17,54	86,95	32,25	Zwischenstand
RL-Checkpoint 62	49,89	68,39	11,78	17,45	73,25	27,31	Zwischenstand
RL-Checkpoint 82	39,84	58,33	10,12	17,74	51,31	26,03	später Zwischenstand
Finaler Checkpoint	37,77	56,26	9,60	17,63	46,67	26,25	Final

Robustheit über Simulationsseeds

Da mikroskopische Verkehrssimulationen stochastischen Einflüssen unterliegen, wurde jede Steuerungsvariante über mehrere Seeds ausgewertet. Die seedübergreifende Auswertung ist maßgebend für die robuste Bewertung der Steuerungsqualität. Sie zeigt nicht nur den mittleren Effekt einer Steuerungsstrategie, sondern auch die Streuung der Ergebnisse. Abbildung 8 zeigt die Streuung der netzweiten Fahrzeugverlustzeit über die 10 Evaluationsläufe. Die Referenzsteuerung weist dabei eine deutlich größere Streubreite auf als die austrainierten RL-

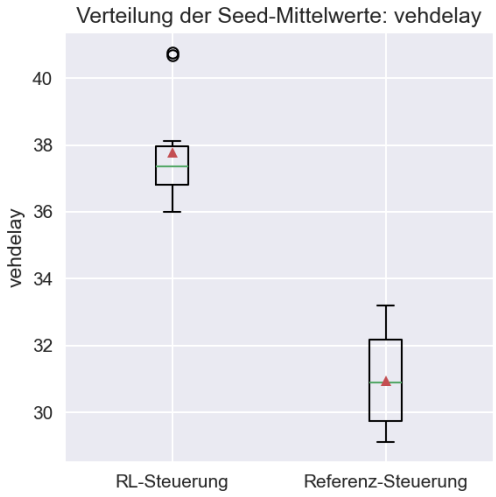


Abbildung 8: Verteilung der netzweiten mittleren Fahrzeugverlustzeit (VEHDELAY) über die 10 Evaluationsläufe je Steuerungsvariante und Trainingsstand. Dargestellt sind Boxplots mit Median, Interquartilsabstand und Ausreißern. Der Mittelwert je Gruppe ist als Dreieck eingezeichnet.

Agenten. Dies deutet darauf hin, dass die gelernte RL-Strategie robust auf unterschiedliche stochastische Verkehrsrealisierungen reagiert und konsistente Steuerungsentscheidungen trifft.

Ergänzende Best-Seed-Auswertung

Neben der seedübergreifenden Bewertung wurde eine ergänzende Best-Seed-Auswertung durchgeführt. Hierbei wurde für jeden betrachteten Trainingsstand derjenige Simulationslauf ausgewählt, der die geringste mittlere Fahrzeugverlustzeit aufwies. Diese Auswertung dient nicht als robuste Hauptbewertung, sondern als Abschätzung des erreichbaren Leistungspotenzials des jeweiligen Checkpoints unter günstiger stochastischer Realisierung. Die

maßgebende Bewertung erfolgt weiterhin anhand der Mittelwerte und Streuungen über alle betrachteten Seeds.

Die Best-Seed-Auswertung wurde insbesondere genutzt, um die verkehrliche Wirkung einzelner Checkpoints detaillierter auf Ebene der LSA und Messstellen zu untersuchen. Dadurch lässt sich erkennen, welche Zufahrten und Knoten im günstigsten Fall besonders stark profitieren und ob verbleibende Schwachstellen auch beim leistungsfähigsten Simulationslauf bestehen bleiben.

Tabelle 8: Ausgewählter Best-Seed je Trainingsstand

Steuerungsvariante/Trainingsstand	Auswahlkriterium	Netzwerk VEHDELAY Best Seed [s/Fzg]	Bemerkung
Referenz	minimaler VEHDELAY	28,74	Vergleichsbasis
RL-Checkpoint 2	minimaler VEHDELAY	103,88	früher RL-Stand
RL-Checkpoint 46	minimaler VEHDELAY	51,77	Zwischenstand
RL-Checkpoint 62	minimaler VEHDELAY	49,89	Zwischenstand
RL-Checkpoint 82	minimaler VEHDELAY	36,27	später RL-Stand
finaler Checkpoint	minimaler VEHDELAY	36,00	finaler RL-Stand

Abbildung 9 zeigt den Best-Seed-Vergleich der netzweiten Kennzahlen. Im günstigsten Simulationslauf erreicht der finale RL-Checkpoint einen VEHDELAY von 36,0 s/Fzg gegenüber 28,7 s/Fzg der Referenzsteuerung (vgl. Tabelle 8). Die Verbesserung gegenüber frühen Trainingsständen ist auch in der Best-Seed-Betrachtung deutlich erkennbar: Iteration 2 weist noch einen VEHDELAY von 103,9 s/Fzg auf, während Iteration 82 bereits auf 36,3 s/Fzg sinkt und damit

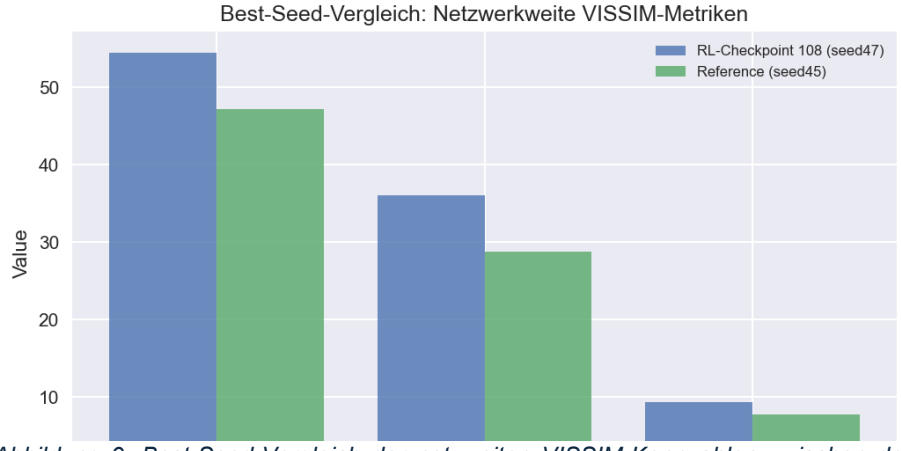


Abbildung 9: Best-Seed-Vergleich der netzweiten VISSIM-Kennzahlen zwischen den evaluierten Trainingsständen und der Referenzsteuerung. Dargestellt sind die Werte des jeweils günstigsten Simulationslaufs (minimaler VEHDELAY) je Steuerungsvariante als gruppierte Balken.

nahe an den finalen Stand heranreicht. Dies bestätigt den in der seedübergreifenden

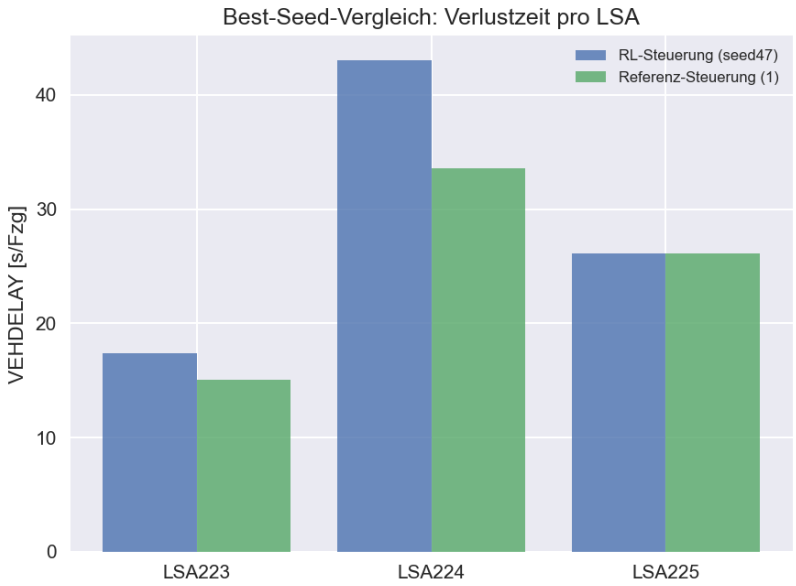


Abbildung 10: Vergleich der mittleren Fahrzeugverlustzeit je LSA zwischen RL-Steuerung (links) und Referenzsteuerung (rechts) im Best-Seed-Vergleich. Die Balkenhöhe entspricht dem VEHDELAY-Mittelwert des jeweils günstigsten Simulationslaufs.

Auswertung beobachteten Lernfortschritt auch unter günstigster stochastischer Realisierung. Die fahrstrombezogene HBS-Auswertung (Tabelle 9 und Tabelle 10) zeigt, dass die RL-Steuerung im Vergleich zur Referenz eine höhere mittlere Wartezeit der B212-Fahrströme aufweist (+28,8 %). Die schlechteste Qualitätsstufe eines B212-Fahrstroms verschlechtert sich von E auf F. Maßgebend hierfür ist der Fahrstrom Columbusstraße Nord, Linksabbieger A1 an LSA 224, dessen mittlere Wartezeit von 94,3 s/Fzg (Referenz, QSV E) auf 110,7 s/Fzg (RL, QSV F) ansteigt. Dieser Fahrstrom stellt die kritischste Zufahrt des gesamten Verbunds dar und bleibt auch unter RL-Steuerung der Engpassfahrstrom. Die Anzahl der B212-Fahrströme mit QSV D–F steigt von 1 auf 2. Demgegenüber zeigen einzelne Fahrströme an LSA 223 und LSA 225 geringfügige Verbesserungen, etwa der Linksabbieger D an LSA 223 (−34 %) und der Geradeausfahrstrom C an LSA 225 (−28,5 %). Die Lloydstraße A Rechts an LSA 225 verbleibt unter RL-Steuerung QSV B.

Fahrstrombezogene HBS-/QSV-Bewertung der B212

Ergänzend zur aggregierten LSA- und Netzbewertung wurde eine fahrstrombezogene HBS-/QSV-Auswertung für die relevanten Ströme der B212 durchgeführt. Hierfür wurden aus den VISSIM-Verlustzeitmessungen je Fahrstrom die Verkehrsstärke, mittlere Wartezeit, mittlere Standzeit, Halte je Fahrzeug sowie die daraus resultierende Qualitätsstufe des Verkehrsablaufs bestimmt.

Tabelle 9: Kompakte HBS-/QSV-Zusammenfassung der B212-Fahrströme

Kennzahl	Referenzsteuerung	austrainierte RL-Agenten
Mittlere Wartezeit B212-Fahrströme [s/Fzg]	22,95	29,57
Schlechteste QSV eines B212-Fahrstroms	E	F
Anzahl B212-Fahrströme mit QSV A–B	15	12
Anzahl B212-Fahrströme mit QSV C	4	6
Anzahl B212-Fahrströme mit QSV D–F	1	2

Die fahrstrombezogene Betrachtung ist erforderlich, da eine Verbesserung der netzweiten Kennzahlen nicht zwangsläufig bedeutet, dass alle maßgebenden Zufahrten der B212 gleichermaßen profitieren. Insbesondere können einzelne Linksabbieger- oder Geradeausströme weiterhin kritisch bleiben, obwohl sich die mittlere Verlustzeit des Gesamtnetzes verbessert. Die QSV-Klassen wurden daher nicht gemittelt, sondern aus den jeweiligen mittleren Wartezeiten je Fahrstrom abgeleitet.

Bereits die Referenzsteuerung erreicht auf den maßgebenden Fahrströmen der B212 überwiegend die Qualitätsstufen A–C; einzelne Ströme der QSV D–F sind im dicht bebauten Innenstadtabschnitt verkehrsplanerisch akzeptabel. Die RL-Steuerung wird somit an einem hohen Ausgangsniveau des Verkehrsablaufs gemessen.

Tabelle 10: Fahrstrombezogene HBS-/QSV-Bewertung der maßgebenden B212-Ströme

LSA	Zufahrt	Signalgruppe	Richtung	Referenz Wartezeit [s/Fzg]	Referenz QSV	RL Wartezeit [s/Fzg]	RL QSV
LSA 223	Columbusstraße Nord	A	Gerade	4,68	A	9,45	A
LSA 223	Columbusstraße Süd	B	Rechts/Gerade	8,08	A	24,65	B
LSA 223	Columbusstraße Süd	D	Links	22,70	B	14,98	A
LSA 223	Kirchenstraße	C	Rechts	27,00	B	21,43	B
LSA 224	Columbusstraße Nord	A	Gerade	22,62	B	27,57	B
LSA 224	Columbusstraße Nord	A	Rechts	21,72	B	23,79	B
LSA 224	Columbusstraße Nord	A1	Links	94,31	E	110,69	F
LSA 224	Columbusstraße Süd	D	Gerade/Links	38,62	C	44,62	C
LSA 224	Columbusstraße Süd	D	Rechts	16,45	A	21,77	B
LSA 224	H.-H.-Meier-Straße	B	Gerade	28,69	B	46,68	C
LSA 224	H.-H.-Meier-Straße	B	Rechts	29,50	B	48,10	C
LSA 224	H.-H.-Meier-Straße	B1	Links	28,83	B	42,05	C
LSA 224	Keilstraße	C	Links	42,39	C	49,11	C
LSA 224	Keilstraße	C	Rechts	44,02	C	52,28	D
LSA 225	Barkhausenstraße	E	Links	39,91	C	43,72	C
LSA 225	Columbusstraße	B	Rechts	11,92	A	12,29	A

LSA 225	Columbus-straße	C	Gerade	30,08	B	21,51	B
LSA 225	Columbus-straße	D	Gerade	23,99	B	25,43	B
LSA 225	Lloydstraße	A	Links	26,43	B	26,39	B
LSA 225	Lloydstraße	A	Rechts	27,07	B	28,17	B

Die fahrstrombezogene Betrachtung (Tabelle 10) bestätigt das heterogene Bild der LSA-Auswertung und gibt zugleich Hinweise auf die Stärken der gelernten Steuerungsstrategie. An LSA 223 zeigen zwei von vier Fahrströmen gegenüber der Referenz verbesserte Wartezeiten: Der Linksabbieger D verbessert sich deutlich von QSV B auf QSV A (−34 %), der Geradeausstrom A verbleibt auf QSV A. Auch an LSA 225 erzielt die RL-Steuerung bei einzelnen Fahrströmen Verbesserungen: Der Geradeausfahrstrom C sinkt von 30,1 s/Fzg auf 21,5 s/Fzg (−28,5 %), die Lloydstraße A Rechts verbleibt stabil in QSV B.

An LSA 224, dem verkehrstechnisch kritischsten Knoten des Verbunds, verbleiben die Herausforderungen erwartungsgemäß am deutlichsten. Der Linksabbieger A1 an der Columbusstraße Nord bleibt der maßgebende Engpassfahrstrom und verschlechtert sich von QSV E auf QSV F (94,3 → 110,7 s/Fzg). Dieser Fahrstrom stellt die kapazitiv anspruchsvollste Zufahrt des Gesamtverbunds dar; dass die RL-Steuerung diesen Engpass nicht auflösen kann, ist konsistent mit dem Befund, dass LSA 224 den stärksten verbleibenden Lernbedarf aufweist und gleichzeitig die größten Lernfortschritte im Trainingsverlauf gezeigt hat.

Insgesamt steigt die Anzahl der B212-Fahrströme mit QSV D–F von 1 auf 2, während die Anzahl der Ströme mit QSV A–B von 15 auf 12 sinkt. Diese Verschiebung ist im Wesentlichen auf LSA 224 zurückzuführen. Die Gesamtstruktur bleibt insofern plausibel, als keine systematische Nichtbedienung erkennbar ist; die Verschlechterungen an einzelnen Fahrströmen zeigen jedoch, dass die Grünzeitverteilung an LSA 224 und 223 noch nicht ausgewogen ist. Die vollständige HBS-/QSV-Auswertung aller Fahrströme ist im Abschlussbericht dokumentiert.

Signalzeitenanalyse

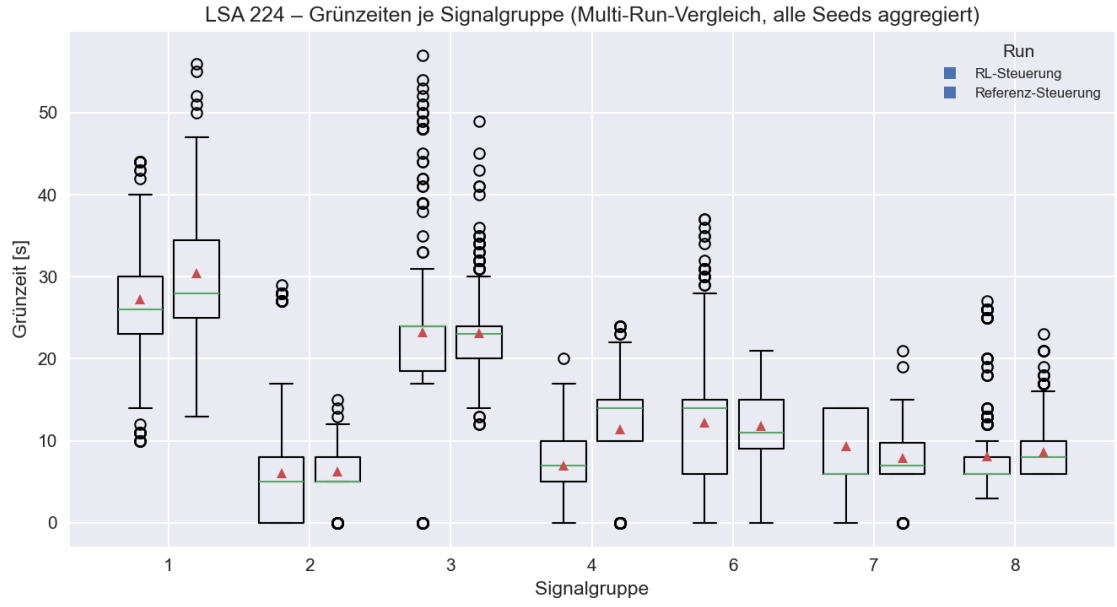


Abbildung 11: Grünzeitverteilungen der Signalgruppen an LSA 224 im Vergleich zwischen RL-Steuerung (linke Boxplots) und Referenzsteuerung (rechte Boxplots). Dargestellt sind Boxplots der expandierten Histogrammverteilungen aus den VISSIM-LZV-Ausgaben, aggregiert über alle Evaluationsseeds.

Zur Plausibilitätsprüfung der gelernten Steuerungsstrategie wurden die Grün- und Rotzeitverteilungen der Signalgruppen ausgewertet. Diese Analyse ist wichtig, da eine Verbesserung der Verlustzeiten nur dann verkehrstechnisch überzeugend ist, wenn die zugrundeliegenden Signalzeitmuster plausibel bleiben. Eine einseitige Bevorzugung einzelner Signalgruppen oder eine systematische Benachteiligung belasteter Phasen würde auf eine unerwünschte Strategie des Agenten hinweisen.

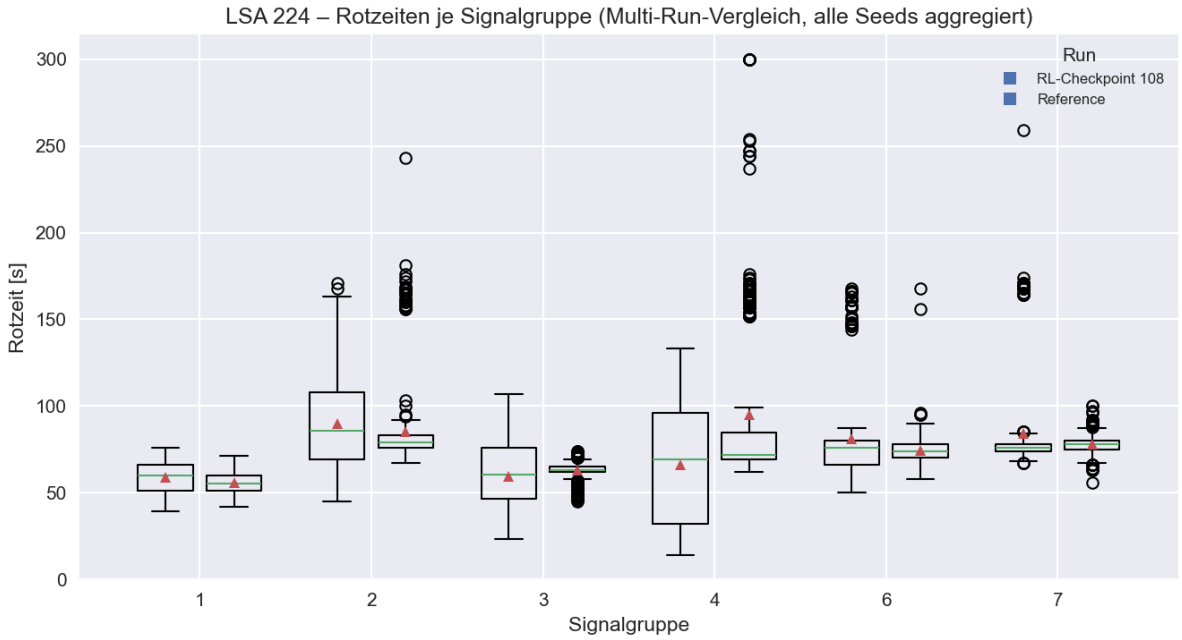


Abbildung 12: Rotzeitverteilungen der Signalgruppen an LSA 224 im Vergleich zwischen RL-Steuerung (linke Boxplots) und Referenzsteuerung (rechte Boxplots). Die Darstellung ermöglicht eine Plausibilitätsprüfung der gelernten Steuerungsstrategie hinsichtlich der Bedienungshäufigkeit einzelner Signalgruppen.

Die Grün- und Rotzeitverteilungen an LSA 224 (Abbildung 11 und Abbildung 12) zeigen, dass die RL-Steuerung im Wesentlichen plausible Signalzeitmuster erzeugt. Die Verteilungen der RL-Steuerung weichen gegenüber der Referenz leicht ab, ohne dass einzelne Signalgruppen systematisch vernachlässigt oder übermäßig bevorzugt werden. Eine vollständige Nichtbedienung einzelner Signalgruppen ist nicht erkennbar. Die verbleibenden Unterschiede in den Grünzeiten einzelner Signalgruppen sind konsistent mit dem beobachteten Lernverhalten: Die RL-Steuerung passt die Phasenzuteilung situationsabhängig an, ohne dabei die Phasenbalance grundlegend zu verletzen.

RL-spezifische Trainings- und Diagnoseauswertung

Die RL-spezifischen Auswertungen werden ergänzend zur verkehrlichen Bewertung herangezogen. Sie dienen nicht als primäres Erfolgskriterium, sondern zur Plausibilisierung des Lernprozesses. Während die VISSIM-Kennzahlen zeigen, ob die Steuerung verkehrlich vorteilhaft ist, geben die RL-Metriken Hinweise darauf, ob der Agent die Reward-Funktion stabil optimiert und ob die gelernten Entscheidungen mit den intendierten Reward-Komponenten zusammenhängen.

Zunächst wurde der Reward-Verlauf des Trainingslaufs betrachtet. Der Reward zeigt, ob das PPO-Training die interne Zielfunktion im Verlauf der Trainingsiterationen verbessert. Zusätzlich wurde der Reward pro Simulationsschritt ausgewertet, um den Lernverlauf unabhängig von möglichen Unterschieden in Episodenlängen besser interpretieren zu können.

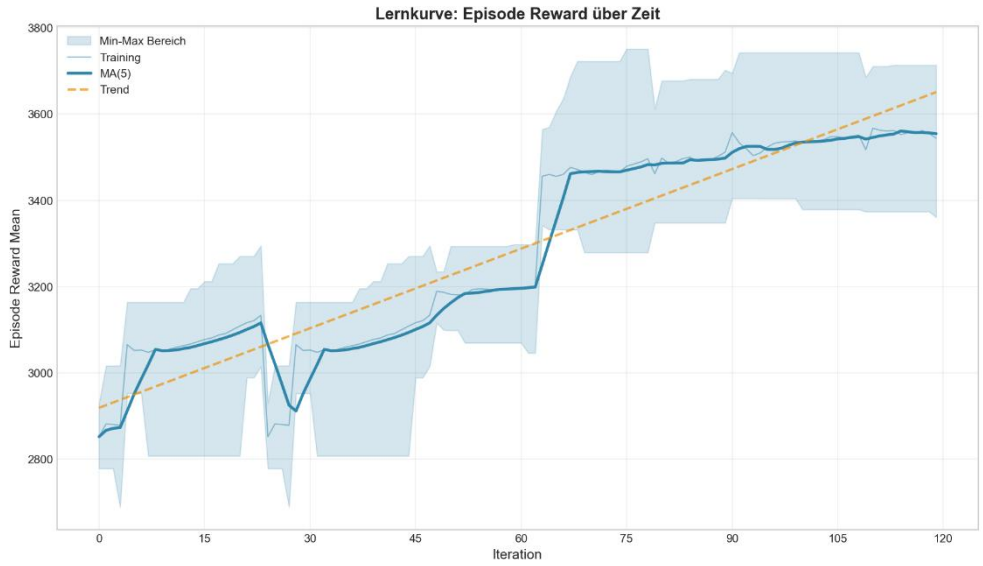


Abbildung 13: Lernkurve des finalen Trainingslaufs. Dargestellt ist der mittlere Episode-Reward über die Trainingsiterationen. Der gleitende Mittelwert (Moving Average – MA) dient der Glättung stochastischer Schwankungen.

Abbildung 13 zeigt den Reward-Verlauf des Trainingslaufs. Der Episode-Reward weist eine klare Aufwärtstendenz auf und stabilisiert sich zum Trainingsende auf einem deutlich höheren Niveau als zu Beginn. Dies bestätigt, dass der PPO-Algorithmus die Reward-Funktion zuverlässig optimiert und das Training keine Instabilitäten oder dauerhaft negativen Lernphasen aufweist. Der gleitende Mittelwert hebt den positiven Grundtrend hervor und zeigt, dass die beobachteten Schwankungen das für stochastisches PPO-Training typische, unkritische Maß nicht übersteigen.

Abbildung 14 zeigt die agentenspezifischen Reward-Komponenten für Agent LSA 224 (vgl. Tabelle 11). Dieser Knoten weist mit einem Gesamt-Reward von 0,41 den niedrigsten Wert aller drei Agenten auf, was direkt mit der topologisch und kapazitiv anspruchsvollsten Situation im Verbund korrespondiert: LSA 224 ist der Knotenpunkt mit den höchsten gemessenen VEHDELAY-Werten und dem maßgebenden Engpassfahrstrom des Gesamtverbunds.

Der hohe Phasen-Strafterm (0,57) zeigt, dass der Agent an LSA 224 dauerhaft mit belasteten Phasen konfrontiert ist, die über längere Zeiträume nicht ausreichend bedient werden können. Dies ist kein Zeichen eines unerwünschten Lernverhaltens, sondern spiegelt die reale Kapazitätsgrenze des Knotens wider: Selbst eine optimale Steuerung kann die strukturell begrenzte Kapazität an LSA 224 nicht vollständig auflösen. Der Fairness-Term (0,31) ist im Vergleich zu LSA 223 (0,07) und LSA 225 (0,11) am höchsten, was auf eine ungleichmäßig verteilte Staubelastung über die Zufahrten hindeutet – konsistent mit dem Befund, dass insbesondere der Linksabbieger A1 an der Columbusstraße Nord als Engpassfahrstrom dominiert.

Positiv hervorzuheben ist, dass der Gesamt-Reward von Agent LSA 224 im Trainingsverlauf eine klare Aufwärtsentwicklung zeigt: Von frühen Trainingsständen mit stark negativen Reward-Werten verbessert sich der Agent systematisch auf das dargestellte Niveau. Dies ist konsistent mit dem in Abbildung 7 gezeigten VEHDELAY-Rückgang von 87,0 s/Fzg (Iteration 46) auf 46,7 s/Fzg beim austrainierten Stand, einem Rückgang um über 46 %. Die Queue-Norm (0,16) und der Fairness-Term sind gegenüber frühen Trainingsständen gesunken, was die

Interpretation stützt, dass Verbesserungen nicht lediglich durch eine Verlagerung von Belastungen zwischen Zufahrten erzielt wurden, sondern auf einer tatsächlichen Reduktion der Rückstausituation beruhen.

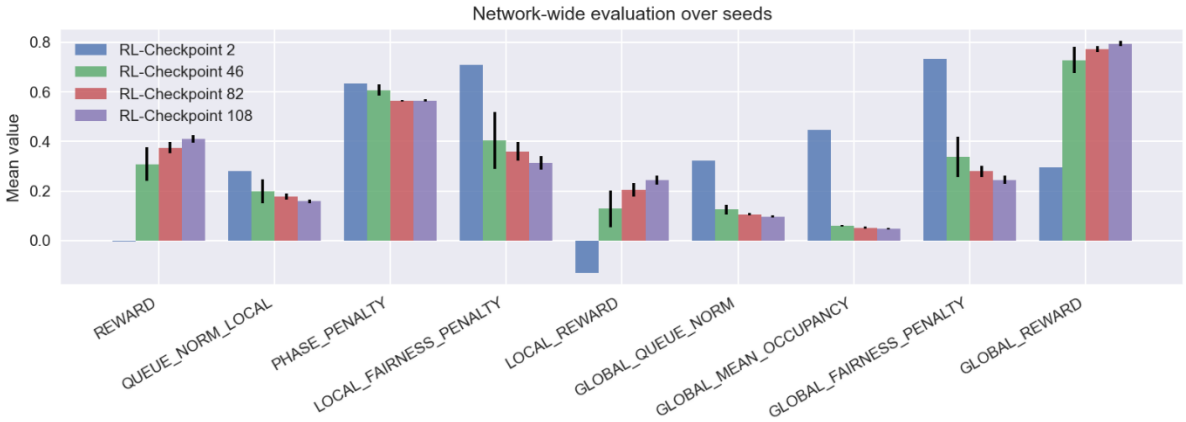


Abbildung 14: Agentenspezifische RL-Diagnosegrößen für Agent LSA 224 im Vergleich der evaluierten Trainingsstände. Dargestellt sind die über alle Evaluationsseeds gemittelten Reward-Komponenten als gruppierte Balken je Trainingsstand. LSA 224 stellt den verkehrstechnisch anspruchsvollsten Knoten des Verbunds dar und weist die kritischsten Reward-Strukturen auf.

Für den Multi-Policy-Ansatz wurden die RL-spezifischen Diagnosegrößen zusätzlich agentenspezifisch ausgewertet. Dadurch kann geprüft werden, ob alle drei LSA-Agenten vergleichbare Lernsignale aufweisen oder ob einzelne Knoten besonders kritisch sind.

Tabelle 11: RL-spezifische Diagnosegrößen je Agent/LSA

Agent / LSA	Reward	Queue-Norm lokal	Phase-Penalty	Fairness-Penalty	Lokaler Reward
LSA 223	0,72	0,04	0,30	0,07	0,68
LSA 224	0,41	0,16	0,57	0,31	0,25
LSA 225	0,77	0,06	0,16	0,11	0,76

Sinkende Queue- und Fairness-Terme deuten darauf hin, dass Verbesserungen nicht nur durch eine Verschiebung von Belastungen zwischen Zufahrten erzielt wurden. Der Phasen-Strafterm zeigt ergänzend, ob belastete Phasen über längere Zeit nicht bedient wurden oder ob die Steuerung eine hinreichend ausgewogene Bedienung der Phasen erreicht.Neben Reward und Reward-Komponenten wurden PPO-spezifische Stabilitätsmetriken betrachtet (siehe Abbildung 15). Dazu gehören Policy Loss, Value Function Loss, Entropy, KL Divergence, Clip Fraction und Explained Variance. Diese Plots werden überwiegend im Abschlussbericht dokumentiert, da sie für die verkehrsplanerische Hauptbewertung nur ergänzenden Charakter haben.

Tabelle 12: PPO-Trainingsdiagnose des finalen Trainingslaufs

Metrik	Beobachtung	Interpretation
Policy Loss	schwankend	Die Schwankungen sind bei PPO erwartbar und wirken im Verlauf unkritisch.

Value Function Loss	Zu Beginn sehr hoch, fällt dann deutlich ab; nach einem erneuten Peak um Iteration ca. 25 sinkt der Loss kontinuierlich auf unter 1.	Der Critic lernt im Verlauf stabilere Wertschätzungen. Die anfänglichen hohen Werte deuten auf eine Einlernphase hin, danach stabilisiert sich die Value Function deutlich.
Entropy	Sinkt kontinuierlich von > 1,5 zu Beginn auf unter 0,5 zum Trainingsende	Erwartetes Verhalten: Der Agent wird mit fortschreitendem Training deterministischer und wählt zunehmend die gelernten, günstigen Steuerungsentscheidungen. Ein verbleibendes Restniveau zeigt, dass eine hinreichende Restunsicherheit für Situationsvarianten erhalten bleibt.
KL Divergence	Verläuft nahezu durchgehend im empfohlenen Bereich 0,01–0,05; vereinzelte kurze Peaks ohne dauerhafte Überschreitung	Die PPO-Clipping-Mechanismen wirken zuverlässig. Keine Hinweise auf dauerhaft aggressive Policy-Updates. Das Training aktualisiert die Policy konservativ und stabil.

Der Episode Reward steigt über die Trainingsiterationen deutlich an und erreicht zum Trainingsende ein höheres Niveau als zu Beginn. Gleichzeitig sinken Value Function Loss, Entropy und KL Divergence, was auf eine zunehmende Stabilisierung des PPO-Trainings hindeutet. Der deutliche Abfall des Reward pro Simulationsschritt ab etwa Iteration 60 fällt mit der Änderung der Episodenlänge von ca. 1.400 auf 2.000 Schritte (bei einem Entscheidungsintervall von 1 s entsprechend Simulationssekunden) zusammen und sollte daher nicht isoliert als Verschlechterung interpretiert werden. Maßgebend für die verkehrliche Bewertung bleiben weiterhin die VISSIM-Kennzahlen; die PPO-Metriken dienen der Plausibilisierung des Lernprozesses.

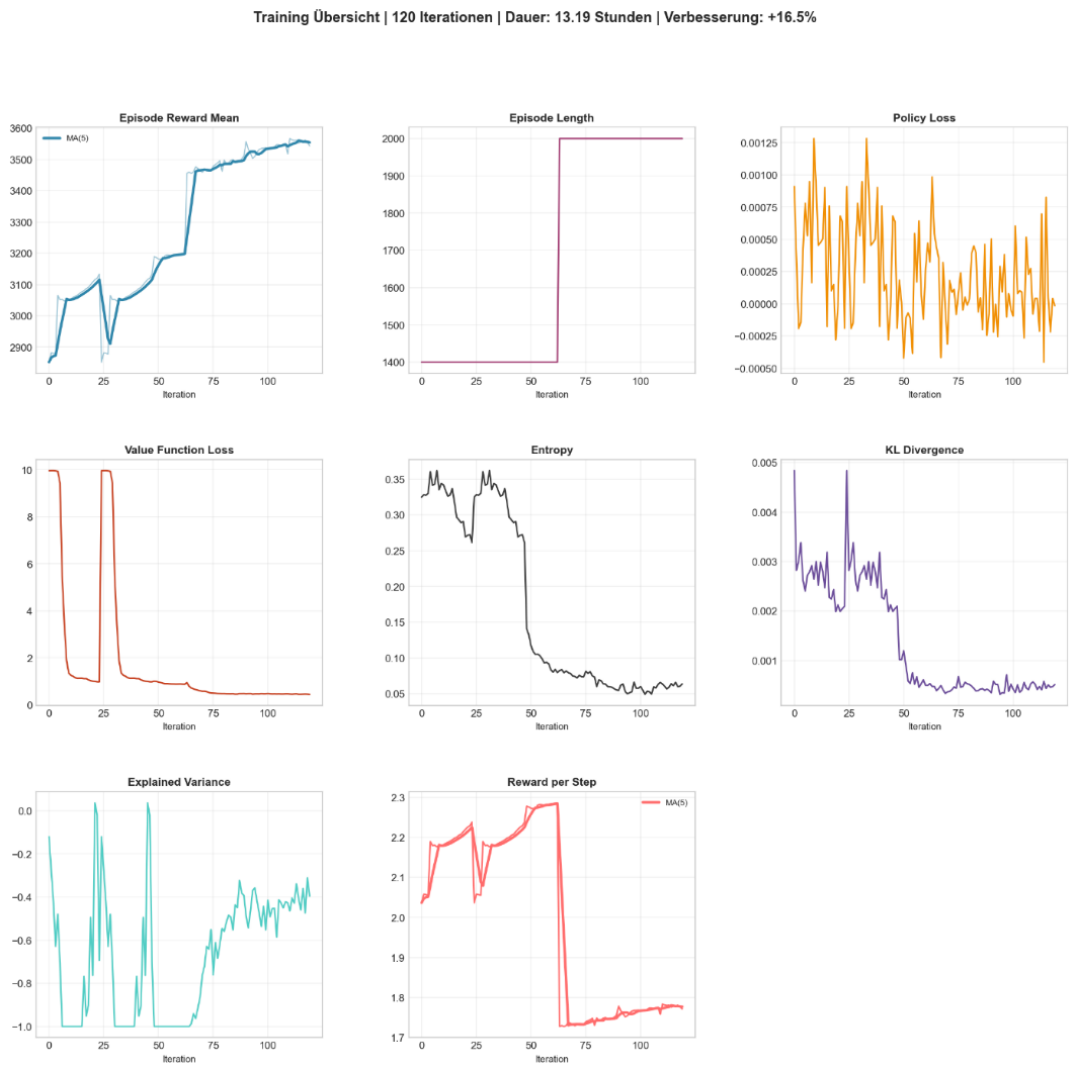


Abbildung 15: PPO-Trainingsdiagnostik des finalen Trainingslaufs. Dargestellt sind die zentralen Stabilitätsmetriken Cumulative Reward Mean, Episode Length, Policy Loss, Value Function Loss, Entropy, KL Divergence, Explained Variance sowie Reward per Step. Die Kurven zeigen den gleitenden Mittelwert über die Trainingsiterationen.

Einordnung der Evaluation

Die Evaluation zeigt die Wirkung der RL-Steuerung aus mehreren Perspektiven. Die klassischen VISSIM-Metriken bewerten die verkehrliche Wirkung gegenüber der Referenzsteuerung. Die Lernverlaufsbewertung über mehrere Checkpoints zeigt, wie sich die verkehrliche Qualität im Trainingsverlauf verbessert. Die LSA- und messstellenspezifischen Auswertungen zeigen, wie sich diese Wirkung räumlich verteilt. Die HBS-/QSV-Bewertung ordnet die Ergebnisse in eine verkehrsplanerisch etablierte Bewertungslogik ein. Die LZV-Auswertung prüft, ob die Signalzeitmuster plausibel bleiben. Die RL-internen Kennzahlen erlauben schließlich eine Interpretation, ob das gelernte Verhalten mit den beabsichtigten Reward-Komponenten übereinstimmt.

Ein wesentliches Merkmal der durchgeführten Evaluation ist ihre unmittelbare Praxisnähe: Training und Evaluation erfolgten vollständig im industriellen Simulationswerkzeug VISSIM, das als De-facto-Standard in der Verkehrsplanungspraxis gilt. Die erzielten Ergebnisse sind damit anschlussfähig an reale Planungsprozesse. Darüber hinaus wurde die bestehende, verkehrstechnisch geprüfte Steuerungslogik der Yunex-Steuergeräte während der gesamten

Evaluation als aktive Sicherheitsschicht integriert: Sie garantiert zu jedem Zeitpunkt die Einhaltung von RiLSA-Konformität, Zwischenzeiten und Mindestfreigabezeiten, ohne dass diese Anforderungen vom neuronalen Netz erlernt werden müssen. Die KI agiert damit ausschließlich innerhalb eines verifizierten, regulatorisch konformen Regelrahmens – eine Voraussetzung, die für einen späteren Feldbetrieb zwingend erfüllt sein muss und die in der akademischen Literatur zu Multi-Agenten-RL-basierter LSA-Steuerung u. E. bislang selten in dieser Form realisiert wurde.

Die Evaluation belegt, dass die entwickelten RL-Agenten im Rahmen des Projekts grundlegende Strukturen einer koordinierten LSA-Verbundsteuerung erlernt haben. Der Trainingsverlauf zeigt eine systematische, deutliche Verbesserung der verkehrlichen Kennzahlen: Der netzweite VEHDELAY sinkt von 58,1 s/Fzg bei Iteration 46 auf 37,8 s/Fzg beim austrainierten Agenten – eine Reduktion um über 35 % allein durch Trainingsdauer, erzielt in einem sequenziellen Training ohne Parallelisierung auf einer einzelnen VISSIM-Instanz.

Einzelne Knoten und Fahrströme werden bereits im aktuellen Trainingsstand auf Referenzniveau oder besser gesteuert: LSA 225 erzielt eine geringfügige Verbesserung gegenüber der Referenz (−1,2 %), einzelne B212-Fahrströme verbessern ihre Wartezeiten, der Linksabbieger D an LSA 223 sogar seine QSV (B → A). Die RL-Agenten weisen zudem eine deutlich geringere stochastische Streuung als die Referenzsteuerung auf, was ihre Robustheit gegenüber wechselnden Verkehrsrealisierungen unterstreicht. Am exemplarisch betrachteten Knoten LSA 224 bleiben die Signalzeitverteilungen plausibel; eine systematische Nichtbedienung belasteter Phasen ist nicht erkennbar.

Der größte verbleibende Abstand zur Referenzsteuerung konzentriert sich auf LSA 224, den topologisch und kapazitiv anspruchsvollsten Knoten des Verbunds. Dass gerade dieser Knoten gleichzeitig den größten Lernfortschritt im Trainingsverlauf aufweist (−46 % zwischen Iteration 46 und dem finalen Stand), bestätigt die grundsätzliche Lernfähigkeit der Agenten auch unter schwierigen Randbedingungen.

Das Niveau der Referenzsteuerung wird mit dem aktuellen Trainingsstand netzweit noch nicht vollständig erreicht. Die Ursachen lassen sich mit dem vorliegenden Design nicht abschließend trennen; begrenzte Trainingsdauer, fehlende Seed-Diversität und die Verbundkomplexität sind plausible Erklärungsfaktoren. Es ist zu beachten, dass die Referenzsteuerung keine einfache oder historisch gewachsene Festzeitsteuerung ist, sondern eine durch BERNARD entwickelte, als „Top-Steuerung“ eingestufte, koordinierte Referenzsteuerung, die nach eigener Einschätzung nahe am verkehrstechnischen Optimum agiert.

Verlängertes Training, eine Erweiterung der Trainings-Seed-Diversität sowie eine gezielte Feinabstimmung der Reward-Gewichte an LSA 224 stellen die vielversprechendsten Ansatzpunkte für eine weitere Annäherung an und perspektivisch eine Überwindung des Referenzniveaus der BERNARD-Optimierung dar. Die im Projekt entwickelte Architektur, bestehend aus VISSIM-basiertem Training, integrierter Sicherheitssteuerung und kooperativer Multi-Policy-Struktur, bildet dafür eine solide und direkt weiterentwickelbare Grundlage.

Über die verkehrlichen Ergebnisse hinaus belegt die Evaluation, dass der gewählte Ansatz grundlegende Voraussetzungen für eine spätere industrielle Anwendung erfüllt: Die RL-Agenten operieren nachweislich jederzeit innerhalb des vorgegebenen Sicherheitsrahmens, erzeugen plausible Signalzeitmuster ohne systematische Phasenausfälle und zeigen eine deutlich geringere stochastische Streuung als die konventionelle Referenzsteuerung, ein relevantes Qualitätsmerkmal für den verlässlichen Betrieb im Feld. Da Training und Evaluation vollständig im industriellen Simulationswerkzeug VISSIM erfolgten, sind die Ergebnisse direkt in bestehende Planungs- und Genehmigungsprozesse übertragbar. Die entwickelte Architektur, bestehend aus VISSIM-basiertem Training, integrierter Sicherheitssteuerung und kooperativer Multi-Policy-Struktur, stellt damit einen nachweisbaren Schritt in Richtung industriell anwendbarer, KI-basierter LSA-Verbundsteuerung dar.

KISLEK zeigt damit, dass sich sichere RL-basierte Verbundsteuerung in einer industriellen VISSIM-Umgebung technisch realisieren, übertragen und weiterentwickeln lässt – ohne

unverhältnismäßig hohe Kosten für bauliche Maßnahmen an den LSA. Die entwickelte Architektur stellt dafür eine skalierbare, simulationsbasierte und industrienahe Grundlage für Forschung, Entwicklung und Umsetzung bereit.

4. Zusammenfassung, Ausblick und Anknüpfungspunkte für Kommunen

KISLEK markiert keinen Endpunkt, sondern einen Übergang: von der Grundlagenforschung an lernbasierten Steuerungsverfahren hin zu einer **anwendungsnahen Entwicklungspfadlogik** für KI in der Verkehrssteuerung.

Die naheliegenden nächsten Schritte sind:

- längere und diversifiziertere Trainingsläufe,
- gezielte Optimierung besonders anspruchsvoller Knoten,
- Erweiterung auf größere LSA-Verbünde,
- Einbindung zusätzlicher Datenquellen, etwa aus V2X-Kontexten,
- Untersuchung alternativer Reward-Strategien, etwa mit stärkerem Fokus auf Emissionen,
- Vorbereitung von Feldtests mit realer Steuergeräteanbindung.

Darüber hinaus zeigt das Projekt, dass die entwickelte Architektur eine belastbare Grundlage für Folgevorhaben bildet. Sie verbindet sichere Steuerungslogik, industrienahe Simulation und lernfähige Multi-Agenten-Strukturen in einer Weise, die sowohl für Forschung als auch für spätere Anwendungsszenarien wertvoll ist.

Auch regulatorisch ist dieser Weg relevant. KI-basierte Steuerungssysteme im Bereich kritischer Infrastruktur werden künftig hohe Anforderungen an Transparenz, Monitoring und Risikomanagement erfüllen müssen. KISLEK adressiert diese Anforderungen bereits implizit durch die Begrenzung des Handlungsraums, die Einbindung bestehender Sicherheitslogik und die nachvollziehbare Evaluationsarchitektur.

Langfristig kann KI-basierte Lichtsignalsteuerung zu einem Werkzeug werden, das Verkehrsplanung und Betrieb sinnvoll ergänzt.

KISLEK ist nicht nur ein Forschungsprojekt zur Steigerung der Leistungsfähigkeit eines Lernalgorithmus. Es ist vor allem ein Projekt zur Erhöhung der **Anschlussfähigkeit von KI in realen Verkehrssteuerungskontexten**. Genau darin liegt sein praktischer Wert.

Erstens zeigt das Projekt, dass KI in der Lichtsignalsteuerung nicht als radikaler Ersatz bestehender Systeme gedacht werden muss. Ein realistischer Weg in die Anwendung ist vielmehr die **Integration in vorhandene Steuerungsarchitekturen**, bei der bestehende Sicherheitslogik erhalten bleibt und die **KI innerhalb klar definierter Schranken optimieren kann**. Das senkt Hürden für spätere Einführungsszenarien erheblich.

Zweitens zeigt KISLEK, dass die technische Machbarkeit nicht nur vom Lernverfahren abhängt. Ein großer Teil der Projektkomplexität lag in **Schnittstellen, Kommunikation, Synchronisierung und Sicherheitsintegration**. Für Kommunen, Hersteller und Integratoren ist das eine zentrale Erkenntnis: Wer KI-basierte Lichtsignalsteuerung einführen will, braucht nicht nur gute Modelle, sondern **belastbare Integrationsarchitekturen**.

Drittens stärkt die vollständige Nutzung von VISSIM die Anschlussfähigkeit an reale Planungs- und Nachweisprozesse. Das Projekt bewegt sich damit nicht nur im Raum algorithmischer Forschung, sondern auch im Raum **verkehrsplanerischer Anwendbarkeit**.

KISLEK hat eine Reihe von Erkenntnissen geliefert, die über den konkreten Untersuchungsraum hinaus relevant sind.

Erstens wurde deutlich, dass **Verbundsteuerung wesentlich komplexer ist als Einzelknotensteuerung**. Die Kopplung benachbarter Knoten erhöht die Anforderungen an Beobachtungsraum, Reward-Design, Trainingsstabilität und Evaluation erheblich. Viele Probleme, die an einer Einzelkreuzung noch überschaubar wirken, entfalten im Verbund neue Dynamiken.

Zweitens hat das Projekt gezeigt, dass **Sicherheit kein nachgelagerter Zusatz**, sondern ein integraler Systembestandteil sein muss. Die Einbindung der bestehenden Steuerlogik war kein Randaspekt, sondern eine zentrale Voraussetzung dafür, dass die Lösung überhaupt in Richtung Praxis gedacht werden kann.

Drittens wurde sichtbar, wie aufwendig **Schnittstellenarbeit** in solchen Projekten tatsächlich ist. Ein erheblicher Teil des Entwicklungsaufwands lag nicht im Entwurf des neuronalen Netzes, sondern in Kommunikation, Synchronisierung und technischer Kopplung zwischen Simulation, KI-Modul und Steuerungslogik.

Viertens machte das Projekt die Spannung zwischen **Praxisnähe und Trainingsgeschwindigkeit** deutlich. Die Nutzung von VISSIM erhöht die Realitätsnähe erheblich, erschwert aber schnelle und breit parallelisierte Trainingsläufe. Für zukünftige Entwicklungen wird es wichtig sein, hierfür methodische und technische Strategien zu finden.

Fünftens zeigen die Ergebnisse klar, dass weiteres Verbesserungspotenzial besteht. Längere Trainingsläufe, größere Seed-Diversität und gezielte Feinabstimmungen, insbesondere am komplexen Knoten LSA 224, sind plausible nächste Schritte, um das Referenzniveau weiter anzunähern oder perspektivisch zu übertreffen.

Für verschiedene Zielgruppen lassen sich daraus konkrete Schlussfolgerungen ableiten:

Tabelle 13: Zielgruppenübersicht.

Zielgruppe	Relevante Erkenntnis
Kommunen	KI kann bestehende LSA-Verbünde perspektivisch ergänzen, ohne Sicherheitslogik aufzugeben und ohne überdimensionierte, kostenintensive Sensorikaufrüstungen im Feld
Verkehrsplanungsbüros	RL kann in bestehende Simulations- und Bewertungsprozesse eingebunden werden
Hersteller und Integratoren	Schnittstellen, Maskierung und Kommunikationslogik sind zentrale Erfolgsfaktoren
Forschung	Verbundsteuerung unter realitätsnahen Randbedingungen bleibt ein hochrelevantes Entwicklungsfeld
Regulierung und Verwaltung	Sichere KI braucht Monitoring, Transparenz und begrenzte Handlungsräume

Die KISLEK-Ergebnisse stellen somit einen großen Schritt Richtung industrieller Anwendung dar. Das Framework bildet die Grundlage für kommunale Pilotprojekte, Produktentwicklung und Feldtests. Wissenschaftlich wurde eine Referenzarchitektur für sicherheitskonformes Multi-Agent-RL geschaffen. Perspektivisch kann die Technologie weniger Stau, kürzere Reisezeiten und weniger Emissionen bei hoher Anpassungsfähigkeit an wechselnde Verkehrsmuster ermöglichen.

Impressum

Projekt:
KISLEK – Künstliche Intelligenz zur Steuerung verbundener Verkehrsknoten

Projektpartner:

- Fraunhofer IOSB-INA
- BERNARD Gruppe ZT GmbH

Assoziierte Partner: PTV Group, BASt, Straßen.NRW, Seestadt Bremerhaven

Ansprechpartner:
Lukas Klein
Fraunhofer IOSB-INA
lukas.klein@iosb-ina.fraunhofer.de

Förderkontext:
gefördert im Rahmen von mFUND
Förderkennzeichen: 19F1185A
Fördermittelgeber: BMV
Projektträger: DLR Projektträger
Laufzeit: 01.06.2024–30.03.2026

Stand: November 2026