KKLab 研究室方針

— AI エージェント経済学（AI Agent Economics）の確立に向けて —

第 2 版

河合勝彦 *

名古屋市立大学大学院経済学研究科

2026 年 5 月 初版 ／ 2026 年 9 月 第 2 版

概要

本稿は，KKLab における研究・教育の方針を「**AI エージェント経済学（AI Agent Economics）**」として整理する．従来の経済学・経営学が「説明する科学」として発展してきたのに対し，本研究室はサイモンの「設計の科学」，ロスの「工学としての経済学」，およびエージェントベース計算経済学の系譜を継ぎ，AI エージェントを媒体として制度・組織・市場を設計・実装・評価・運用する研究室として位置づけられる．AI エージェントは，(a) 経済主体のモデル，(b) 制度の参加者，(c) 設計・運用の道具という三つの視座から扱われる．研究は，モデル化・実装・実験／評価・運用の四工程を循環させ，妥当性検証の階梯に沿って結果の信頼性を高める．構成員には，「エージェントを設定する力」から「エージェントシステムを設計・実装・評価・運用する力」を経て，最終的に「AI エージェント経済学者として独立する力」へと至る三段階の発達経路を提示する．第 2 版では，学問的背景に LLM エージェント研究の到達点を加え，重点研究テーマ，妥当性検証の原則，研究室の運営と実践規範，想定される批判への応答，用語集を新設した．本方針は研究室活動を通じて継続的に改訂される．

キーワード: AI エージェント経済学，設計の科学，工学としての経済学，メカニズムデザイン，マーケットデザイン，エージェントベース計算経済学，計算社会科学，LLM エージェント，シリコンサンプリング，研究室方針

目次

* 連絡先: kkawai@econ.nagoya-cu.ac.jp

1. ビジョン

1.1 研究室の使命

KKLab は，AI エージェントを媒体として，経済・経営・組織・制度を「説明する」だけでなく「**設計する**」科学を実践する研究室である．本研究室はこの実践知を **AI エージェント経済学（AI Agent Economics）**と位置づける．

研究室の使命を一文で述べれば，「望ましい制度・組織・市場を，AI エージェントを用いて構想し，動くかたちで示し，検証し，現実に届ける」ことである．ここで「動くかたちで示す」とは，理論モデルを紙の上で閉じるのではなく，相互作用する主体のシステムとして実装し，その振る舞いを観察可能・比較可能・改善可能にすることを意味する．

1.2 なぜ「工学」ではなく「経済学」と呼ぶのか

ここで意図的に「工学」ではなく「経済学」と呼ぶ理由は三つある．第一に，研究対象は依然として人間，組織，市場，制度などの社会経済システムであり，その因果構造，インセンティブ，戦略的相互作用といった経済学的問いが分析の中心に位置するからである．第二に，AI エージェントは設計対象であると同時に，経済主体のモデルでもあり，その振る舞いは経済学の理論で読み解かれ，また経済学の理論を駆動する設計道具にもなるからである．第三に，設計の良否を判断する価値基準——効率性，公正性，誘因整合性，厚生——は経済学が長年にわたり洗練させてきたものであり，技術的に「動く」ことと経済的に「望ましい」こととを区別する視点を与えるからである．すなわち，AI エージェント経済学は，経済学を「**工学的に**」実践する一形態である．

1.3 定義

定義（AI エージェント経済学） AI エージェント経済学とは，AI エージェントを (a) 経済主体のモデル，(b) 制度・市場・組織の参加者，(c) 設計と運用の道具として用い，社会経済システムを説明・設計・実装・評価・運用する，経済学の実践的分野である．

この定義は，AI エージェントを単なる分析ツールとしても，単なる研究対象としても扱わない点に特徴がある．エージェントは，モデルであり，主体であり，道具である．どの側面が前景に出るかは研究課題によって異なるが，三つの側面が相互に依存していることを常に意識する．

1.4 三つの視座

定義に含まれる三つの視座を表 1 と図 1 に整理する．

三つの視座は排他的ではなく，一つの研究の中で移り変わる．たとえば混雑料金制度の研究では，まず住民の行動モデルとしてエージェントを用い（対象），次に料金設定アルゴリズムそのものをエージェントとして市場に置き（参加者），最後に制度パラメータの探索と評価を LLM と推論手法に委ねる（道具）．一つの研究課題を三つの視座で言い換えられるかどうかは，その課題が AI エージェント経済学の課題として熟しているかどうかの目安になる．

視座	エージェントの位置づけ	典型的な問い	主な方法
対象	人間・組織の行動モデル（ホモ・シリクス）	人々はこの制度のもとでどう振る舞うか	合成被験者実験，ABM，シリコンサンプリング
参加者	市場・組織の一員として振る舞う主体	エージェント同士，エージェントと人間が相互作用する市場は何を生むか	マルチエージェント・シミュレーション，ゲーム理論，学習動学
道具	設計者・分析者・運用者の補助	どの制度が望ましく，どう探索・検証・運用するか	自動メカニズムデザイン，LLM アブダクション，シミュレーションベース推論

表 1　AI エージェント経済学の三つの視座

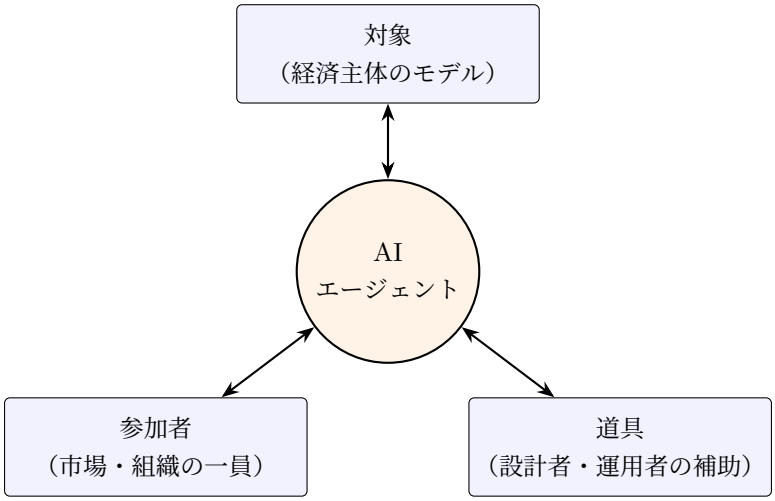

図 1　AI エージェントの三つの位置づけ

2. 学問的背景

2.1 説明する科学としての経済学・経営学

伝統的な経済学・経営学は，企業，組織，市場，人間の意思決定がどのような原理で動いているのかを明らかにする「**説明する科学**（explanatory science）」として発展してきた．均衡分析，計量経済学，組織論，戦略論のいずれも，観察された現象の構造を理解することを主目的としてきた．

「説明」と「設計」は対立するものではない．よい設計は正確な説明に依存し，設計の失敗はしばしば説明の欠落を露わにする．エプスタインが「生成できないなら，説明したことにならない（If you didn’t grow it, you didn’t explain it）」と述べたように，現象を生み出すメカニズムを構築して動かすことは，それ自体が説明の一形態でもある．本研究室は，説明する科学の蓄積を土台としつつ，そこに設計・実装・運用の工程を接続する．

2.2 サイモンの設計の科学

ハーバート・サイモンは『人工物の科学』（邦訳『システムの科学』）において，専門職の知的活動の中心には「設計」があると論じた．サイモンにとって設計とは，現在の状況を望ましい状況へ変えるための行為の構想であり，工学だけでなく経営，医療，教育，政策にも共通する知的活動である．

サイモンは，専門職教育において分析だけでなく「**設計の科学**（science of design）」を確立する必要性を主張した．

サイモンの設計論は，同じく彼が提唱した**限定合理性**（bounded rationality）と**満足化**（satisficing）の概念と不可分である．人間の認知資源は有限であり，最適化ではなく「十分によい」解を探索する．この見方は，LLM エージェントの振る舞いを理解するうえでも示唆的である．LLM エージェントは完全合理的な最適化主体ではなく，与えられた文脈と手がかりに依存して行動する限定合理的主体として振る舞う．したがって，AI エージェント経済学は，ホモ・エコノミクスの仮定を緩めた経済学——行動経済学，組織の経済学——と自然に接続する．

サイモンはまた，人工物を「内部環境」と「外部環境」との接面として捉えた．制度や組織という人工物の性能は，それ自体の構造だけでなく，参加者の認知と環境の性質との適合によって決まる．エージェントシステムの設計においてこの視点は，エージェントの設計（内部）と制度・市場の設計（外部）を切り離さずに考える必要性として現れる．

2.3 ロスの「工学としての経済学」

アルヴィン・ロスは「**工学としての経済学**（economist as engineer）」という見地から，経済学者が市場を分析するだけでなく市場を設計する役割を担うようになったことを指摘した．マーケットデザインには，単純な理論モデルだけでなく，制度の細部，参加者の行動，実験，計算，運用上の制約を扱う工学的アプローチが必要であり，ゲーム理論，実験経済学，計算経済学はそのための相補的な道具である．

ロス自身の仕事——研修医マッチング制度の再設計，腎移植交換，学校選択制——は，安定マッチングの理論だけでは制度は動かず，参加者の戦略的行動，制度の細部，運用上の制約，制度の歴史が結果を左右することを示した．同様にミルグロムは，周波数オークションの設計において，複雑な制約のもとで価格を「発見」する仕組みを計算的手法と結びつけた．市場は自然物ではなく人工物であり，設計と保守の対象である，という認識がマーケットデザインの出発点である．

2.4 エージェントベース計算経済学の系譜

AI エージェント経済学のもう一つの源流は，**エージェントベース計算経済学**（Agent-based Computational Economics: ACE）である．テスファッションは ACE を，相互作用する自律的エージェントの動的システムとして表現された経済過程の計算的研究と定義し，均衡を前提とせず，異質な主体の局所的相互作用から集計的パターンが創発する過程をボトムアップに扱う枠組みを整備した．エプスタインとアクステルの『人工社会の成長』（Sugarscape モデル）は，交易，文化，人口動態が単純なルールから生成されることを示した古典である．近年の総説（Axtell and Farmer, 2025）は，エージェントベースモデル（ABM）が経済学・ファイナンスにおいて，異質性・非均衡動学・制度の細部を扱う手段として成熟してきたことを整理している．

ACE の限界も知られている．エージェントの行動ルールを研究者が手で書かねばならず，その恣意性が結果の説得力を損ないやすいこと，そしてパラメータの校正と推定が困難なことである．LLM エージェントは前者を，シミュレーションベース推論（SBI）は後者を緩和する可能性を持つ．本研究室は ACE の方法論的蓄積——実験計画，感度分析，モデル文書化——をそのまま継承する．

2.5 LLM エージェントの登場と「ホモ・シリクス」

大規模言語モデル（LLM）の登場は，エージェントの行動ルールを手で書く代わりに，自然言語で状況を与えて人間らしい応答を生成させることを可能にした．ホートン（Horton, 2023）は LLM を「**ホモ・シリクス**（homo silicus）」――人間の行動を模した計算的主体――と呼び，社会的選好，価格設定，最低賃金に関する古典的実験を LLM で再現できることを示した．アーガイルら（Argyle et al., 2023）は，LLM に人口統計的属性を与えて人間標本を模倣する「**シリコンサンプリング**」を提案し，世論調査の回答分布の一部を再現できることを示した．アーら（Aher et al., 2023）は古典的な心理学・行動経済学実験の複製を「チューリング実験」として体系化し，パークら（Park et al., 2023）の「生成的エージェント」は，記憶・省察・計画を備えたエージェント集団が社会的相互作用を創発させることを示した．チェンら（Chen et al., 2023）は GPT の選択行動が顕示選好の観点から高い合理性を示すことを報告し，マニングら（Manning et al., 2024）は LLM に仮説生成から実験設計・実行・分析までを担わせる「自動化された社会科学」の構想を示した．マクロ経済の領域でも，LLM エージェントによる景気動態の再現（Li et al., 2024）が試みられている．

一方で慎重な結果も蓄積している．ビスビーら（Bisbee et al., 2024）は，LLM の合成回答が人間データに比べて分散が過小で，モデル更新に対して不安定であることを示し，サントゥルカーら（Santurkar et al., 2023）は LLM の「意見」が特定の集団に偏ることを示した．これらは，LLM エージェントを人間の代替として無条件に使うことへの警告であり，妥当性検証（第 3.3 節）を研究方針の中核に置く根拠である．

もう一つの重要な潮流は，AI エージェントが**市場の参加者**となる状況の研究である．カルヴァーノら（Calvano et al., 2020）は，Q 学習に基づく価格設定アルゴリズムが明示的な通信なしに暗黙的共謀を学習することを示し，フィッシュら（Fish et al., 2024）は LLM ベースの価格設定エージェントでも超競争的価格が生じ得ることを報告した．パークスとウェルマン（Parkes and Wellman, 2015）は，合理性の理想像としての「マキナ・エコノミクス（machina economicus）」を論じ，AI 主体からなる経済を分析する経済学の必要性を早くから指摘している．さらに，深層学習によるオークション設計（Dütting et al., 2019），強化学習による税制設計（Zheng et al., 2022）など，制度設計そのものを計算的に探索する研究も進んでいる．

2.6 AI エージェントによる統合

AI エージェントの登場は，サイモンの「設計の科学」とロスの「工学としての経済学」を，実装可能な形で結びつける．エージェントは環境を認識し，目標に基づいて行動し，他のエージェントや人間と相互作用する経済主体として設計できる．これにより，メカニズムデザインやマーケットデザインは，抽象的な理論にとどまらず，実際に動作し，評価され，改善されるエージェントシステムとして実装されていく．理論モデル，シミュレーション，実験，運用データを往復させながら，望ましい制度や市場を継続的に改善することが可能になる．

表 2 は，本研究室が各系譜から何を継承するかをまとめたものである．

この意味で，AI エージェント経済学は，単に AI を使って業務を自動化する技術ではない．それは，人間，組織，市場，制度を含む複雑な社会システムを，**説明し，設計し，実装し，評価し，運用する**ための新しい実践知である．

系譜	本研究室が継承するもの
説明する科学（経済学・経営学）	因果構造・誘因・戦略的相互作用の理論，計量的検証の作法，評価の価値基準
サイモンの設計の科学	設計を知的活動の中心に置く姿勢，限定合理性・満足化，内部環境と外部環境の接面
ロスの工学としての経済学	制度の細部と運用を扱う工学的態度，理論・実験・計算の併用
エージェントベース計算経済学	ボトムアップの生成的説明，異質性と非均衡動学，感度分析と文書化の作法
LLM エージェント研究	自然言語で駆動する行動モデル，合成被験者，自動化された社会科学，妥当性への警戒
アルゴリズム的ゲーム理論・機械学習	計算的メカニズム設計，学習する主体の均衡分析，アルゴリズム的共謀の理解

表 2　学問的系譜と本研究室の継承

3. 研究方針

3.1 研究循環の四工程

KKLab における研究は，以下の四工程を循環させる（図 2）.

1. **モデル化**：扱う制度・市場・組織を，エージェント，効用，戦略，情報構造，相互作用ルールとして定式化する.
2. **実装**：エージェントフレームワーク（ADK, OpenAI Agents SDK, Claude Agent SDK, Mesa, 独自実装等）を用いて，モデルを動作するエージェントシステムとして実装する.
3. **実験・評価**：シミュレーション，人間参加型実験，A/B テスト，運用データの分析により，設計された制度の振る舞いを検証する.
4. **運用と改善**：実装された制度・サービスを実環境で運用し，得られた知見を理論モデルに還元する.

この四工程は線形ではなく循環として運用される．理論的に望ましい制度が現実の運用で機能しないことは珍しくなく，運用上の発見はしばしば新しい理論的問いを生む.

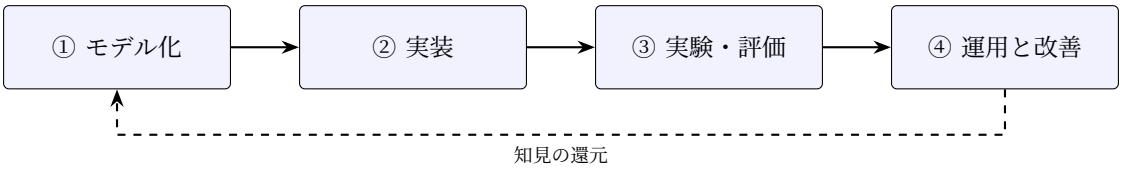

図 2　KKLab の研究循環

3.2 各工程の要点

モデル化．モデル化で最初に決めるのは，何を固定し，何を設計変数とするかである．エージェント（属性，情報，目的，行動の自由度），制度（ルール，メッセージ空間，配分と支払い），環境（外生ショック，時間構造，ネットワーク），評価指標の四つを明示的に書き分ける．理論モデルが存在す

る領域では，まず理論の予測を書き下し，シミュレーションが再現すべき基準として用いる．

実装．実装は特定のフレームワークに依存しない．ただし，(i) 乱数シード・温度・モデルのバージョンを固定して再実行可能にすること，(ii) 全てのプロンプト，応答，ツール呼び出しをログとして保存すること，(iii) 小さな構成で単体テストを通してから規模を拡大すること，(iv) API 費用と計算時間を事前に見積もることを必須とする．「動いた」と「再現できる」は別であり，後者のみが研究成果として扱われる．

実験・評価．評価は複数の基準で行う（表 3）．単一指標の最適化は，他の基準の劣化を見落とす．制度比較では，同一の乱数・同一のエージェント集団を用いた対照条件を設け，差の推定に統計的手続きを適用する．人間参加型実験（oTree 等）や A/B テストと組み合わせる場合には，事前登録と倫理審査を行う．

基準	問い	代表的な指標
効率性	資源は望ましく配分されているか	総余剰，配分効率，厚生損失
公正性	負担と便益の分配は正当化できるか	分配の格差指標，手続き的公正の評価
誘因整合性	正直な行動が各主体にとって得か	逸脱による利得，戦略的操作の検出
頑健性	前提の摂動に対して結論は保たれるか	パラメータ・モデル・プロンプトの感度分析
安定性	動学は収束するか，振動・崩壊しないか	収束時間，価格・配分の分散
安全性	暴走・搾取・誤情報が生じていないか	逸脱行動の頻度，異常検知
コスト	現実的な資源で運用できるか	計算時間，API 費用，人手

表 3　制度・エージェントシステムの評価基準

運用と改善．運用では，監視指標を定め，人間が介入できる仕組み（human-in-the-loop）と巻き戻し手順をあらかじめ用意する．運用で得られた観察は，モデル化の前提の修正，評価指標の追加，新しい理論的問いのいずれかとして研究循環に戻す．

3.3 妥当性検証の原則

LLM エージェントの結果は，それ単体では証拠にならない．人間データ，理論，別モデルとの三角測量によってはじめて証拠になる．本研究室は，妥当性を表 4 の階梯で段階的に確認する．

水準	名称	確認すること
0	顔妥当性	出力が課題の文脈として意味をなし，明らかな破綻がない
1	理論的・様式化された事実の再現	既知の理論結果（需給均衡，最後通牒ゲームの提案分布など）が再現される
2	人間データとの定量的突合	人間実験・調査の分布や効果量と統計的に整合する
3	予測的妥当性	新しい条件のもとでの人間の結果を，事前に予測できる
4	運用上の外的妥当性	実環境での運用データが，設計時の予測と整合する

表 4　妥当性検証の階梯

階梯とあわせて，次の手続きを標準とする．

- **感度分析**：モデル，プロンプトの言い回し，温度，乱数，ペルソナ表現を変えて結論が保たれる

か確認する．

- **複数モデルでの反復**：単一ベンダーのモデルに依存した結論を避け，オープンウェイトモデルを含む複数モデルで再実行する．
- **事前登録**：主要な仮説，条件，分析手順を実行前に記録し，事後の選択的報告を防ぐ．
- **訓練データ汚染の点検**：古典的実験の複製では，モデルがその実験の結果を「知っている」可能性を考慮し，新規課題や変形課題で確認する．
- **否定的結果の報告**：再現に失敗した条件も，LLM エージェントの適用範囲を画定する知見として記録・公開する．

3.4 重点研究テーマ

三つの視座と研究循環を踏まえ，本研究室は当面，次の六つのテーマを重点とする．

T1 合成被験者と社会科学実験の自動化． LLM を合成被験者として用いる際の妥当性の条件を明らかにし，人間データによる校正（キャリブレーション）と偏りの補正手法を開発する．問い：どの種類の課題・集団において，シリコンサンプリングは人間標本の代替または補完となり得るか．手法：古典的実験の複製，人間調査との突合，複数モデル比較．

T2 マルチエージェント市場シミュレーションと制度比較． オークション，マッチング，価格設定，取引プラットフォームなどの制度を，LLM エージェントと古典的 ABM の双方で実装し，制度間の性能を比較する．問い：アルゴリズム的価格設定主体が多数を占める市場で共謀・不安定性はどの条件で生じるか，どの制度がそれを抑止するか．手法：マルチエージェント・シミュレーション，学習動学，ゲーム理論的分析．

T3 メカニズム・マーケットデザインの計算的探索． LLM による仮説生成（アブダクション），ABM による定式化，シミュレーションベース推論による逆推論，メカニズムデザインによる選択を一つのパイプラインとして運用する．問い：望ましい結果を生む制度パラメータや制度構造を，計算的にどこまで探索できるか．手法：自動メカニズムデザイン，SBI，多目的最適化．

T4 組織とエージェントチームの経済学． 複数の AI エージェントと人間からなる組織における分業，委任，誘因，情報共有を，組織の経済学とプリンシパル・エージェント理論の枠組みで分析・設計する．問い：エージェントチームの生産性と信頼性は，どのような役割分担・報酬設計・監督構造のもとで高まるか．手法：組織シミュレーション，契約理論，運用ログの分析．

T5 エージェント経済の制度設計． AI エージェントが消費者・企業に代わって取引・交渉・購買を行う「エージェント経済」において必要となる制度——本人確認，委任と責任，決済，消費者保護，API を介した企業間取引——を構想し，設計上の要件を導く．問い：エージェント同士の取引が主流化した市場では，どの既存制度が機能を失い，どの新制度が必要になるか．手法：概念的分析，制度シミュレーション，事例研究．

T6 教育への応用：エージェントを作りながら学ぶ経済学． 均衡分析，古典的 ABM，LLM エージェントの三つの視点を往復させながら経済学を学ぶ教材を開発し，研究室の教育に組み込む．問い：「動くエージェントを書いて確かめる」学習は，経済学の理解と設計能力の獲得にどう寄与するか．手法：教材開発，授業実践，学習効果の測定．

3.5 方法論的立場：説明・推論・設計の接続

本研究室の方法論は，説明する科学と設計する科学の往復を，推論の種類として明示することを重視する．

- **アブダクション（最良の説明への推論）**：観察された現象を生み出し得るメカニズム仮説を複数生成する．LLM は仮説候補の大量生成と批評に用いる．
- **演繹（シミュレーション）**：仮説を ABM やエージェントシステムとして定式化し，前提から帰結を導いて重要パラメータを特定する．
- **帰納（推定・検証）**：人間データや運用データを用いて，パラメータを推定し（SBI，計量経済学），仮説を検証する（因果推論）．
- **設計（規範的選択）**：制約・倫理・実行可能性の観点から制度の候補を比較し，選択する（メカニズムデザイン）．

この四つは研究循環の四工程と対応しており，それぞれの推論がどの段階でなされ，どの証拠に依拠するかを研究計画の段階で明示する．説明の妥当性を担保するのは帰納と因果推論であり，設計の妥当性を担保するのは複数基準による評価と運用での検証である．

4. 学習目標：三段階の発達経路

KKLab の構成員には，次の三段階の発達経路を提示する（表 5）．

4.1 段階 1：エージェントを設定する力

LLM，プロンプト，ツール，メモリ，評価指標などを用いて，目的に沿ったエージェントを「設定」できること．プロフェッショナルな AI エージェントビルダーとしての基本能力に対応する．到達の目安は，目的に沿ったエージェントを一体作り，評価セットで性能を測り，改善の記録を残せることである．典型的な成果物は，経済実験の被験者エージェントや調査回答エージェントとその評価レポートである．

4.2 段階 2：エージェントシステムを設計・実装・評価・運用する力

複数エージェントによる相互作用，制度設計，市場メカニズムを含むシステムを，エージェントフレームワークを用いて設計・実装・評価・運用できること．AI エージェントエンジニアとしての中核能力に対応する．到達の目安は，制度ルールを持つ市場・組織のシミュレーションを実装し，条件比較の実験を設計・実行し，再現パッケージ付きで報告できることである．典型的な成果物は，オークション形式の比較や混雑料金制度のシミュレーション研究である．

4.3 段階 3：AI エージェント経済学者として独立する力

経済学・経営学の理論を踏まえ，現実の制度設計・組織設計・市場設計の問いを立て，エージェントシステムを通じて解き，学術的・実務的な貢献をなすこと．本研究室が育成する最終像である．到達の目安は，先行研究に位置づけた研究課題を設定し，理論・実装・実験を統合した論文を書き，実

務・政策への含意を示せることである．典型的な成果物は，修士論文，査読付き論文，共同研究の報告書である．

段階	役割像	中核能力	到達の目安（典型的な成果物）
1	AI エージェントビルダー	エージェント単体の設定・調整・評価	評価セット付きのエージェント一体と改善記録
2	AI エージェントエンジニア	マルチエージェント／制度の設計・実装・運用	制度比較シミュレーションと再現パッケージ
3	AI エージェント経済学者	制度・組織・市場設計の問いを立て解き貢献する	理論・実装・実験を統合した論文と政策・実務含意

表 5　KKLab の三段階発達経路

段階間の移行は線形ではない．段階 3 の研究課題に取り組む過程で段階 1 の基礎に戻ることは通常であり，段階 1 にある構成員も段階 3 の問いに触れることで，何のために技術を学ぶのかを理解する．学部生には段階 1 を，修士課程では段階 2 を経て段階 3 に至ることを標準的な経路として想定する．

5. 必要なスキルセット

上記の発達経路を支えるため，KKLab では次の領域を体系的に学習する．

- **経済学・経営学の理論**：ミクロ経済学，ゲーム理論，メカニズムデザイン，マーケットデザイン，組織の経済学，戦略論，行動経済学．
- **計算と実装**：Python，LLM ／エージェントフレームワーク，ABM フレームワーク（Mesa 等），分散システム，データ基盤，再現性のある実験環境．
- **実験と統計**：実験デザイン，因果推論，ベイズ統計，シミュレーションベース推論，計量経済学，A/B テスト運用．
- **設計と運用**：プロダクト設計，制度設計，サービス運用，リスク管理，倫理審査．
- **研究の作法**：文献探索と批判的読解，研究計画の作成，学術的文章，再現パッケージの作成，発表．

表 6 は，各領域について段階ごとに期待する水準を示す．

領域	段階 1	段階 2	段階 3
理論	主要概念の理解	制度をモデルとして定式化	先行研究への位置づけと理論的貢献
計算と実装	単体エージェントの構築	マルチエージェント・制度の実装と運用	研究基盤の設計と公開
実験と統計	評価指標の設計と測定	条件比較実験と統計的推論	因果推論・SBI による推定と妥当性検証
設計と運用	要件の言語化	監視・介入を含む運用設計	制度の社会実装と政策含意
研究の作法	記録と報告	再現パッケージと発表	論文執筆と査読対応

表 6　領域別・段階別に期待する水準

これらは独立した科目ではなく，エージェントシステムを実際に作って動かすという一つの実践のなかで統合的に学ばれる．学習の順序は，理論を先に完成させてから実装に進むのではなく，小さな実装を通じて理論の必要性を実感し，理論を学んで実装を改善する往復を基本とする．

6. 研究成果と発表方針

KKLab の研究成果は次の三層で整理される．

1. **学術成果**：査読付き論文，ワーキングペーパー，コンファレンス発表．
2. **実装成果**：オープンソースで公開されるエージェントシステム，ベンチマーク，データセット．
3. **実践成果**：企業・公共セクターとの共同設計プロジェクト，政策提言，教育教材．

これら三層は独立ではなく，同一の研究プロジェクトから順に派生していくことを期待する．すなわち，理論的問いから始まった研究が実装され，現場で運用され，最後に再び論文として理論に還元される，という往復が標準的な成果形態である．

公開にあたっては次を原則とする．

- **早期公開**：ワーキングペーパーはプレプリントサーバやリポジトリ（Zenodo 等）で DOI を付して早期に公開し，査読を待たずに議論に供する．
- **再現パッケージの同梱**：論文には，コード，プロンプト，設定，ログ，分析スクリプトを含む再現パッケージを添付する．
- **成果物の単位**：研究プロジェクトごとに一つのリポジトリを設け，論文・コード・データを同じ場所で管理する．
- **発表先の選択**：経済学・経営学の学会に加え，計算社会科学，マルチエージェントシステム，経済学と計算の接点にある国内外の学会・ワークショップを積極的に用いる．

7. 研究室の運営と実践規範

7.1 研究の進め方

研究は，研究計画書（問い，先行研究，方法，評価，スケジュール）の作成から始まる．週次の研究会で進捗を共有し，中間報告と最終報告を節目とする．各構成員は研究ノートを継続的に記録し，失敗した試みも含めて経緯を残す．研究ノートは，再現パッケージと論文の「方法」節の原資料になる．

7.2 コードと文書の管理

研究プロジェクトは GitHub 上のリポジトリを単位として管理する．各リポジトリには README，ビルド手順，依存関係の固定（バージョン指定），実行手順を置く．LLM を用いる研究では，使用したモデル名とバージョン，プロンプト，温度などの設定，および全ての入出力ログを保存し，モデルの更新によって結果が変わり得ることを前提に記録する．論文は LaTeX で執筆し，ソースと PDF をともにリポジトリで管理する．

7.3 生成 AI・エージェントの研究利用の原則

本研究室は，生成 AI とエージェントを研究の協働者として積極的に用いる．コーディング，文献探索，文章の推敲，仮説生成などに AI を利用することは奨励される．ただし，次の原則を守る．

- **責任は人間にある**：AI が生成した内容を成果物に含める場合，その正しさと妥当性の責任は構成員が負う．
- **検証を省略しない**：AI が生成したコードは実行とテストにより，引用は原典の確認により，主張は根拠の確認により検証する．存在しない文献の引用は重大な研究不正として扱う．
- **利用を開示する**：論文・報告書では，生成 AI をどの工程にどのように利用したかを記載する．
- **自律性の水準を意識する**：AI に委ねる範囲（提案のみ／人間の確認付き実行／自律実行）を作業ごとに明示し，研究上の判断は人間が行う．
- **データの扱い**：個人情報や未公開データを外部の AI サービスに送信する際は，契約と倫理審査の条件に従う．

7.4 共同研究と外部連携

企業・自治体・他研究機関との共同研究は，研究循環の「運用と改善」の工程を実現するために重要である．共同研究では，研究の問い，データの扱い，成果の公開範囲，知的財産の帰属を事前に合意し，文書化する．研究室として公開を原則としつつ，相手方の事情に応じて公開の時期と範囲を調整する．

8. 倫理と行動規範

AI エージェントが経済主体として振る舞う以上，設計者の責任は重い．KKLab では次を行動規範とする．

- **再現可能性**：コード，データ，実験条件を可能な限り公開する．
- **透明性**：エージェントの設計意図，評価指標，限界を明示する．
- **公正性**：制度設計が特定主体に過度な負担を強いないか継続的に点検する．
- **安全性**：自律的に動作するエージェントについて，暴走，誤情報，悪用に対する設計上の対策を必須とする．具体的には，権限の最小化，人間による停止・介入の手段，実環境への影響範囲の限定を設計段階で組み込む．
- **人間参加型研究の倫理**：人間を被験者とする実験・調査は，所属機関の倫理審査を経て実施し，インフォームド・コンセントと個人情報保護を徹底する．LLM 生成のペルソナが実在の個人を模倣しないよう配慮する．
- **評価の誠実性**：都合のよい条件・モデル・乱数の選択的報告を行わない．失敗した条件も記録し，適用範囲の限界として報告する．
- **社会的責任**：研究対象としての制度・市場が，現実の人々の生活に影響することを忘れない．アルゴリズム的共謀や搾取的な設計のように，技術的に可能でも社会的に望ましくない設計については，その危険性を明らかにする側に立つ．

9. 想定される批判への応答

AI エージェント経済学に対しては，いくつかの根本的な批判が予想される．本研究室はそれらを退けるのではなく，研究方針の中に応答として組み込む．

批判 1：LLM は人間ではない．その通りである．しかし，ホモ・エコノミクスもまた人間ではなく，有用なモデルであった．問題は「人間と同じか」ではなく「どの問いについて，どの程度，人間の代理として有用か」であり，これは経験的に決まる．本研究室は妥当性の階梯（第 3.3 節）を用い，LLM エージェントの結果を人間データ・理論・別モデルとの三角測量によって位置づける．人間の代替ではなく，仮説生成と制度の事前検証の道具として用いることを基本とする．

批判 2：閉じたモデルに依存した研究は再現できない．商用モデルは更新され，同じプロンプトが同じ結果を返す保証はない．本研究室は，モデルのバージョン固定と全ログの保存，オープンウェイトモデルを含む複数モデルでの反復，モデル非依存の結論のみを主張する慎重さをもってこれに応える．モデルの更新によって結果が変わること自体も，LLM エージェントの性質に関する知見として記録する．

批判 3：それは工学であって学問ではない．ロスが示したように，制度を設計し運用する営みは，理論では見えなかった問いを生み，理論を前進させてきた．設計に関する知——何がなぜ機能し，機能しないか——は蓄積可能で一般化可能な知識である．本研究室は，問いを先に立て，設計と実装をその問いに答える手段として用いることで，工学的実践を学問的貢献につなげる．

批判 4：古典的実験の複製は訓練データの記憶にすぎない．この危険は現実であり，複製実験の成功だけでは LLM エージェントの妥当性を主張しない．新規課題，変形課題，実施後に公開された人間実験との比較によって，記憶ではない一般化能力を確認する．また，複製実験は妥当性の階梯の一段にすぎず，それ自体を研究の目的とはしない．

批判 5：エージェント経済は投機的な想定である．アルゴリズム的価格設定はすでに多くの市場で運用され，AI エージェントによる購買・交渉・取引の仕組みも実装が進んでいる．制度が事後的に固定される前に，その設計原理を検討しておくことが，まさに設計の科学の役割である．

10. 結語

説明する科学と設計する科学が統合される地点に，AI エージェント経済学の大きな可能性がある．サイモンが構想した設計の科学，ロスが実践した工学としての経済学，ACE が整備したボトムアップの方法論，そして LLM エージェントがもたらした行動モデルの新しい表現力は，いま初めて一つの研究実践のなかで結びつけられる．KKLab は，この新しい実践知を，教育・研究・実装・運用のすべての層において育てていく．本方針はその出発点であり，研究室の活動を通じて継続的に改訂されていく文書である．

付録 A. 用語集

用語	説明
AI エージェント	環境を認識し，目標に基づいて行動を選択し，ツールや他の主体と相互作用する計算的主体．本稿では LLM を中核とするものを主に指す．
LLM エージェント	大規模言語モデルを意思決定の中核に置き，自然言語による状況記述から行動を生成するエージェント．
マルチエージェントシステム	複数の自律的エージェントが相互作用するシステム．市場・組織のシミュレーションの基本形．
エージェントベースモデル（ABM）	異質な主体の局所的相互作用から集計的現象を生成するシミュレーションモデル．経済学での応用が ACE．
ホモ・シリクス	人間の行動を模した計算的主体としての LLM を指すホートンの用語．
シリコンサンプリング	LLM に人口統計的属性等を与え，人間標本の回答分布を模倣させる手法．
合成被験者	人間の被験者の代わりに，または補完として実験・調査に用いる LLM エージェント．
メカニズムデザイン	望ましい結果を導くようにルール（メカニズム）を設計する理論．誘因整合性が中心概念．
マーケットデザイン	メカニズムデザインを実際の市場・制度の設計と運用に適用する実践的分野．
誘因整合性	各主体にとって正直に行動することが最適であるような性質．
アルゴリズム的共謀	価格設定アルゴリズム同士が明示的な合意なしに超競争的価格に到達する現象．
シミュレーションベース推論（SBI）	尤度が明示的に書けないシミュレーションモデルについて，観測データからパラメータの事後分布を推定する手法群．
アブダクション	観察を最もよく説明する仮説を推論すること（最良の説明への推論）．
人間参加型実験	人間の被験者が制度やエージェントと相互作用する実験．oTree 等のプラットフォームで実施．
Human-in-the-loop	自律的システムの動作に人間の確認・介入を組み込む設計．
再現パッケージ	研究結果を第三者が再現するために必要なコード，データ，設定，ログ，手順の一式．
妥当性の階梯	LLM エージェントの結果の信頼性を，顔妥当性から運用上の外的妥当性まで段階的に確認する枠組み（第 3.3 節）．

付録 B．研究室関連文書

本方針と対をなす研究室内文書として，次のものがある（いずれも研究室リポジトリで管理）．

- 『設計する科学の研究手法——LLM アブダクション × ABM × SBI × メカニズムデザイン』：

第 3.5 節の方法論的立場を展開した論文.

- 『設計する科学としての経営戦略・マーケティング戦略』：上記の方法論を経営戦略・マーケティングに適用した姉妹編.
- 『エージェントを作りながら学ぶ経済学』：均衡分析・古典的 ABM・LLM エージェントの三視点を往復する教科書（テーマ T6）.
- 『API エコノミーの進化と Deep Agents』：エージェント経済の制度的背景を扱う論文（テーマ T5）.
- 『Mesa への LLM 統合』ガイドブック：ABM フレームワークへの LLM エージェント統合の技術解説（テーマ T2・T3 の実装基盤）.
- 『社会科学系の AI コーディング三段階モデル』：第 7 節の「自律性の水準」の考え方の出典.

改訂履歴

- **2026 年 5 月 初版**：ビジョン，学問的背景（説明する科学・サイモン・ロス），研究循環の四工程，三段階の発達経路，スキルセット，成果の三層，倫理規範，結語.
- **2026 年 9 月 第 2 版**：定義と三つの視座を追加．学問的背景に ACE と LLM エージェント研究の系譜を追加．各工程の要点，評価基準，妥当性検証の階梯，重点研究テーマ（T1–T6），方法論的立場を新設．発達経路に到達の目安を追加．研究室の運営と実践規範，想定される批判への応答，用語集，研究室関連文書を新設．参考文献を拡充.

主要参考文献

設計の科学と工学としての経済学

- Simon, H. A. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69(1), 99–118.
- Simon, H. A. (1996). *The Sciences of the Artificial* (3rd ed.). MIT Press.（邦訳：稲葉元吉・吉原英樹訳『システムの科学 第 3 版』パーソナルメディア，1999 年）
- March, J. G., & Simon, H. A. (1958). *Organizations*. Wiley.
- Roth, A. E. (2002). The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics. *Econometrica*, 70(4), 1341–1378.
- Roth, A. E. (2015). *Who Gets What—and Why: The New Economics of Matchmaking and Market Design*. Houghton Mifflin Harcourt.
- Milgrom, P. (2004). *Putting Auction Theory to Work*. Cambridge University Press.
- Milgrom, P. (2017). *Discovering Prices: Auction Design in Markets with Complex Constraints*. Columbia University Press.
- Hurwicz, L., & Reiter, S. (2006). *Designing Economic Mechanisms*. Cambridge University Press.
- Milgrom, P., & Roberts, J. (1992). *Economics, Organization and Management*. Prentice Hall.

エージェントベース計算経済学

- Epstein, J. M., & Axtell, R. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press / MIT Press.
- Epstein, J. M. (2006). *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press.
- Tesfatsion, L., & Judd, K. L. (Eds.). (2006). *Handbook of Computational Economics, Vol. 2: Agent-Based Computational Economics*. North-Holland.
- Axtell, R. L., & Farmer, J. D. (2025). Agent-Based Modeling in Economics and Finance: Past, Present, and Future. *Journal of Economic Literature*, 63(1), 197–287.

LLM エージェントと経済学・社会科学

- Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? NBER Working Paper No. 31122.
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), 337–351.
- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR 202.
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.
- Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The Emergence of Economic Rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51), e2316205120.
- Manning, B. S., Zhu, K., & Horton, J. J. (2024). Automated Social Science: Language Models as Scientist and Subjects. NBER Working Paper No. 32381.
- Li, N., Gao, C., Li, M., Li, Y., & Liao, Q. (2024). EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4), 401–416.
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR 202.
- Immorlica, N., Lucier, B., & Slivkins, A. (2024). Generative AI as Economic Agents. *ACM SIGecom Exchanges*, 22(1).

アルゴリズム・機械学習とメカニズムデザイン

- Parkes, D. C., & Wellman, M. P. (2015). Economic Reasoning and Artificial Intelligence. *Science*, 349(6245), 267–272.
- Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review*, 110(10), 3267–3297.
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806.
- Dütting, P., Feng, Z., Narasimhan, H., Parkes, D. C., & Ravindranath, S. S. (2019). Optimal Auctions through Deep Learning. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, PMLR 97.
- Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI Economist: Taxation Policy Design via Two-Level Deep Multiagent Reinforcement Learning. *Science Advances*, 8(18), eabk2607.
- Nisan, N., Roughgarden, T., Tardos, É., & Vazirani, V. V. (Eds.). (2007). *Algorithmic Game Theory*. Cambridge University Press.
- Wooldridge, M. (2009). *An Introduction to MultiAgent Systems* (2nd ed.). Wiley.

邦語文献

- 坂井豊貴（2013）『マーケットデザイン——最先端の実用的な経済学』ちくま新書.
- アルビン・E・ロス（2016）『Who Gets What——マッチメイキングとマーケットデザイン』櫻井祐子訳，日本経済新聞出版社.
- ハーバート・A・サイモン（1999）『システムの科学 第 3 版』稲葉元吉・吉原英樹訳，パーソナルメディア.