Using LLMs to Implement Normative Reasoning Capabilities of Autonomous Agents

SUPERVISORS

Dr. Nisansa de Silva / Dr. Surangika Ranathunga
Prof. Stephen Cranefield / Prof. Tony Savarimuthu (Otago, NZ)

PRESENTED BY
TEAM AUTOBOTS

Navindu - 200110P
Prabhash - 200144X
Kavindu - 200694G

September 4, 2024

Agenda

1	Introduction
2	Problem Description
3	Research Objectives
4	Related Works
5	Methodology
6	Testing & Evaluation
7	Feasibility
8	Timeline

Introduction

Autonomous Agents in Multi-Agent Societies

- An **Autonomous agent** is a system situated within an environment that can perceive and perform actions to accomplish a set of objectives.*
- When performing tasks, it is expected that these agents behave according to human **norms** and standards set by society
    - Especially in social scenarios when interacting with humans and even other AI agents
- This leads to the concept of a **multi-agent system**

* - Wang et al., “A Survey on Large Language Model based Autonomous Agents” (2024)

Introduction

The Rise of LLM Agents

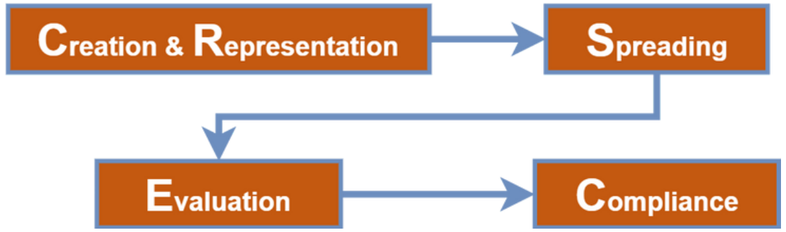
- **LLM**s began to be used in autonomous agents as the core computational engine to plan and memorize interactions
  - These agents are known as **LLM Agents***
  - Providing the ability to process human norms in natural language instead of condition rules
  - With the ability to interact as naturally as a human would

* - Leng et al., “Do LLM Agents Exhibit Social Behavior?” (2024)

Introduction

Learning the Norms

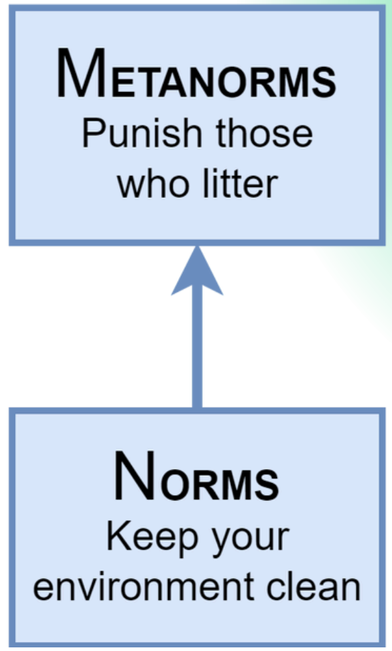
- Providing all social rules for each agent can be impractical depending on the environment complexity and interactions between agents.
- Through **norm emergence**, we expect the agents to learn the social norms through the behaviours and guidance of other agents, similar to human society
- *CRSEC** is one such norm emergence framework used to model the steps utilized to handle different personas within multi-agent LLM societies



* - Ren et al., “Emergence of Social Norms in Generative Agent Societies: Principles and Architecture” (2024)

Introduction

Meta norms - The Next Level of Norms



- When agents are spreading the norms, it is important to
  - Ensure agents comply with norms
  - Wrongful actions are penalized
- For this purpose, higher order norms - known as **metanorms*** show means of achieving through either,
  - Implementing a punishment to the agent that violated a norm OR
  - Behaving differently to agents with low reputations due to norm violation

* - Axelrod, “An Evolutionary Approach to Norms” (1986)

Problem Statement

- Recently, LLM agents posing as human personas, were able to build and spread social norms to other agents without human input.
    - However, there exists a **research gap** in normative agents where they **lack a dynamic, natural language framework to punish non-cooperative agent behaviour** (such as norms violation or non-punishment)
    - Causing **slow norm adoption by non-cooperative agents**

Problem Statement

Motivation for Research Problem

- Natural language would **uncover valuable explanations for understanding multi-agent behaviour**
  - Critical in today's context to explain AI
- An effective norm/metanorm mechanism would **discourage the spread of non-cooperative behaviours**
- **Reduce the interference by defective agents** when performing tasks
- Ability to **adapt dynamically to novel interaction scenarios** and limit reliance on rigid rules and conditions.

Research Objectives

Main Objective

Implement an approach to model the emergence of punishment-based metanorms through natural language in multi-agent systems

i

Facilitate
the spread of cooperative behaviours through the norm dynamics from human social scenarios

ii

Extend
the current Smallville simulation environment to monitor the behaviour of the LLM agents in various social dilemma scenarios

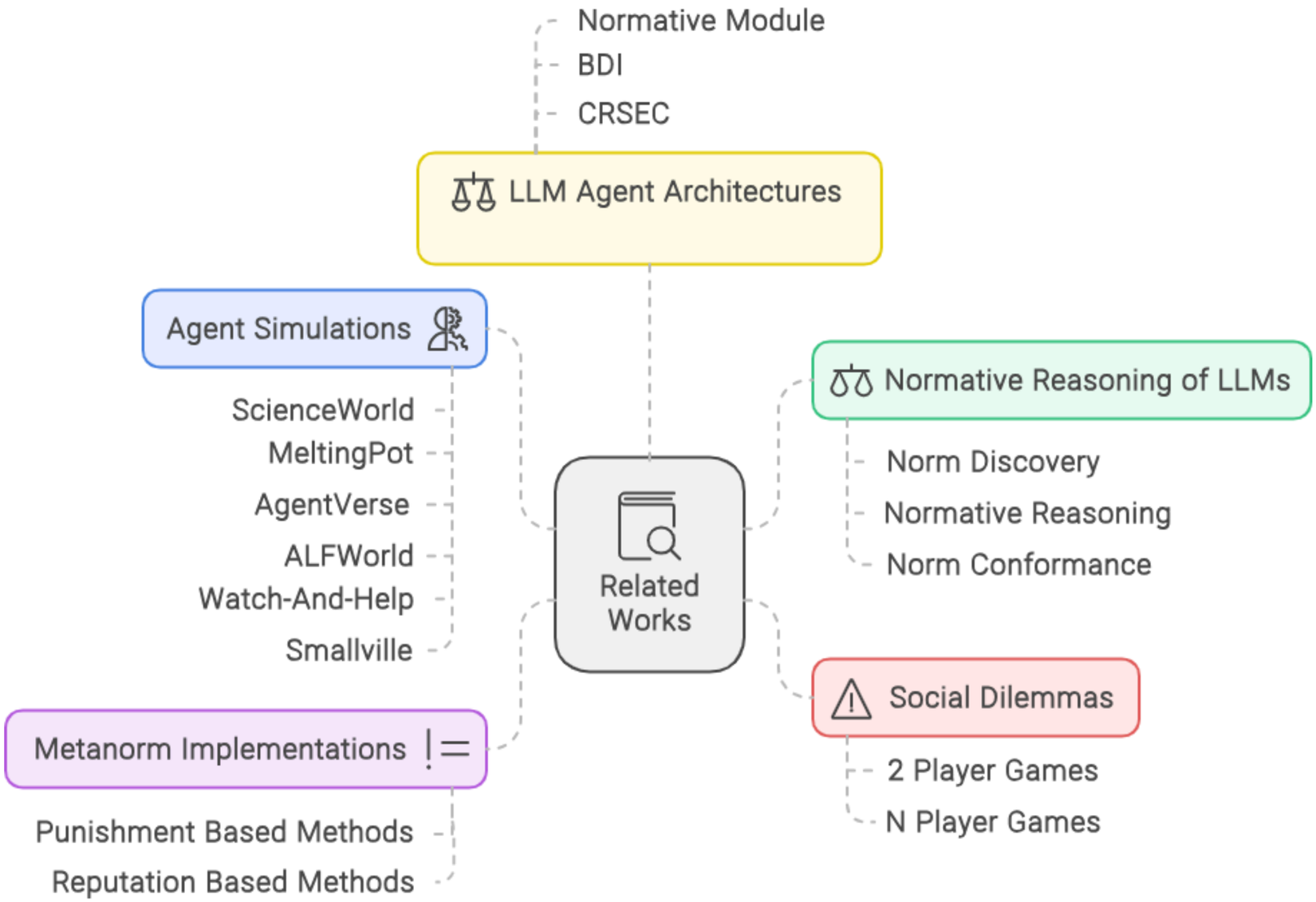
iii

Assess
the effectiveness of the normative capabilities of LLM agents within our proposed implementation using human evaluation methods

Metanorm spread

Related Works

Mind-Map of Reviewed Literature



Related Works

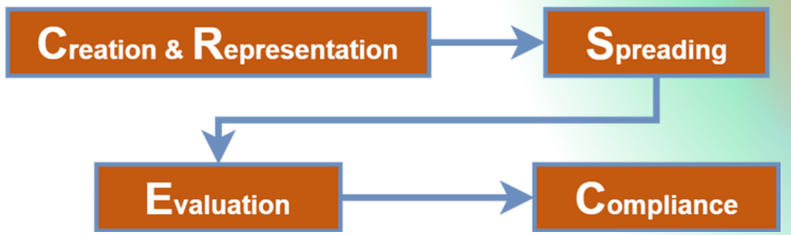
AI Agent Simulations

	Social?	LLMs?	Visual?	Multi-agent?
ScienceWorld [1]	✗	✗	✗	✗
MeltingPot [2]	✓	✗	✓	✓
AgentVerse [3]	✗	✓	✓	✓
ALFWorld [4], **Watch & Help** [5]	✗	✗	✓	✓
Smallville [6]	✓	✓	✓	✓

We see the Smallville [6] as the most appropriate to simulate multi-agent LLM systems

Related Works

Normative Architectures for LLMs

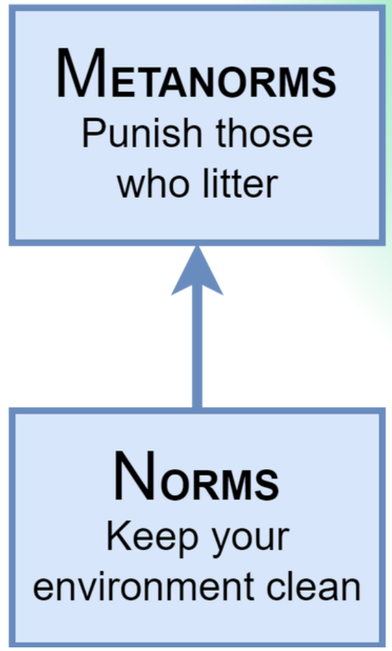


- A norm-capable agent should be able to perform
  - Norm discovery, normative reasoning & norm conformance [7]
- The complete norm emergence process within LLM agents can be represented by the CRSEC architecture [8]
- Other examples → BDI (Belief, Desire, Intention) model [9], Normative LLM model [10]
  - But they don't capture the internal process of developing norms

Creation & Representation	The initial setup of agent memory
Spreading	Norm Discovery
Evaluation	Norm Reasoning
Compliance	Norm Conformance

Related Works

Metanorms Applications



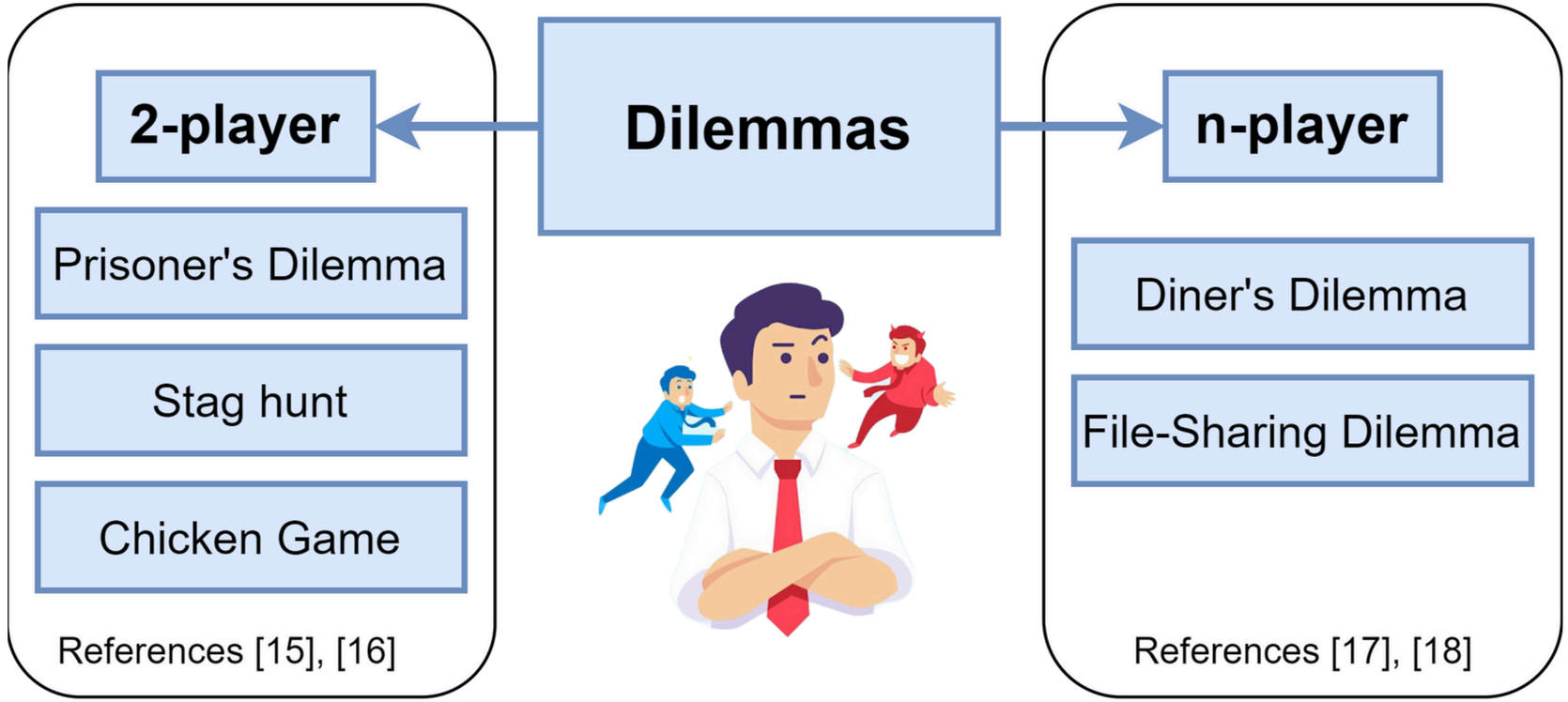
- Higher order norm for agents to respond to norm violations
- Enforce through,
  - Sanction / Punishment based [11, 12]
  - Indirect Reciprocity / Reputation based [13]

- Reputation-based metanorm framework would require a large knowledge base and memory
- Punishment-based metanorms reduce two types of defection,
  i. Direct norm violations
  ii. Non-punishment of violators

Related Works

Social Dilemmas and Strategies

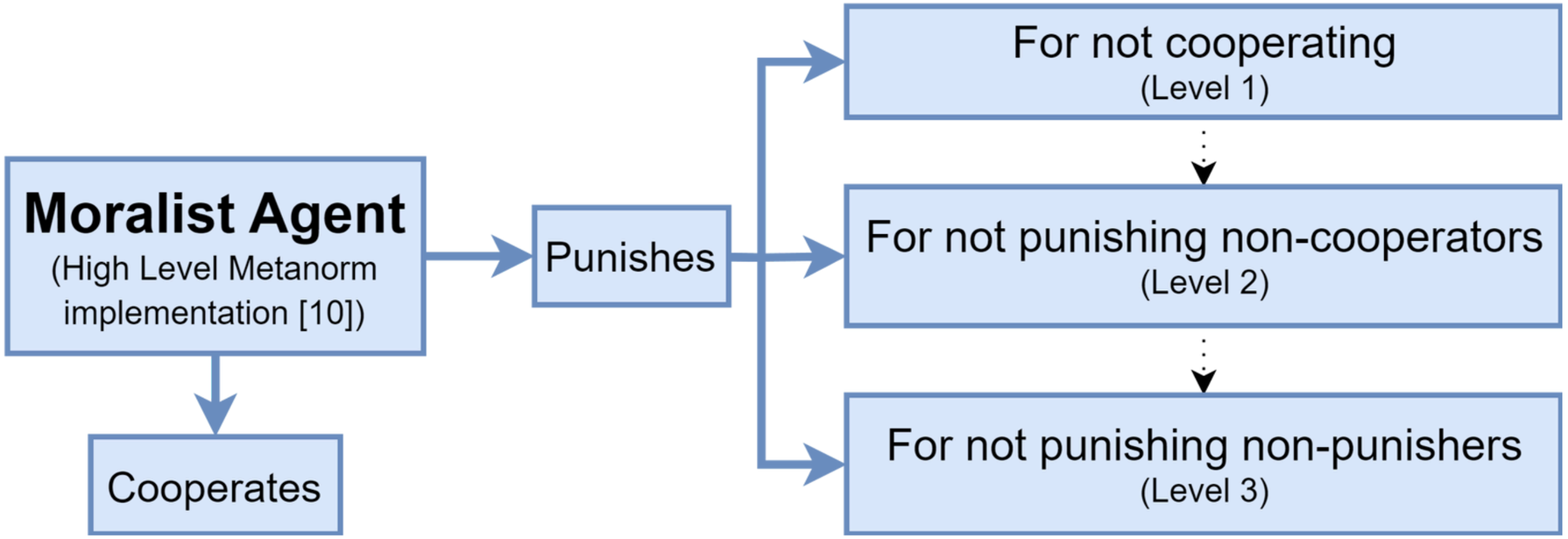
- Agents face a dilemma [14]
  - Maximize personal utility in the short term → **Defect**
  - Cooperate with others for long-term benefits → **Corporate**



Related Works

Social Dilemmas and Strategies

- For an n-player system to be maintained, there should be a sufficient number of **“Moralist Agents”** in the system [12]

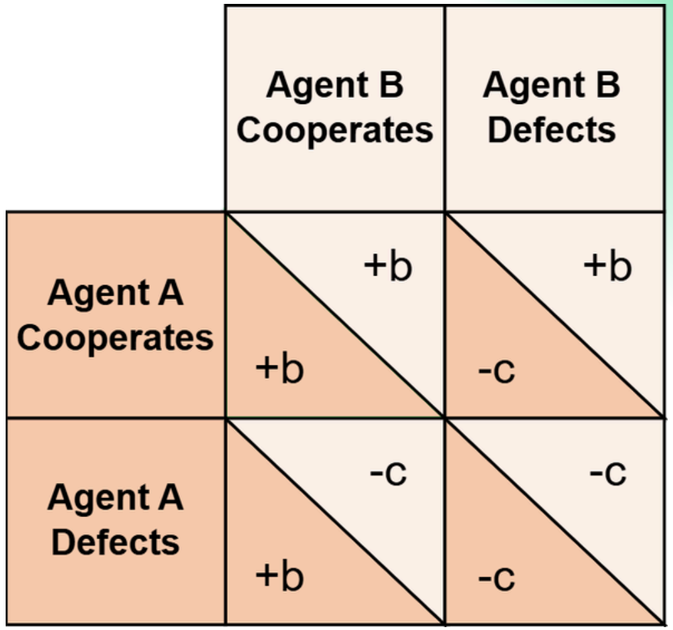


- Ensures agents with defective strategies are controlled from consuming the population

Related Works

Social Dilemmas + Metanorms

- Metanorms have been extensively studied and stimulated with social dilemmas
- Utilizes the pay-off matrices along with scores for punishing to help agents select appropriate actions → leading to cooperation
- However, current metanorm implementations are **limited to mathematical constraints or conditional rules**
- And most simulations consider abstract hypothetical scenarios



Payoff matrix for 2-player game

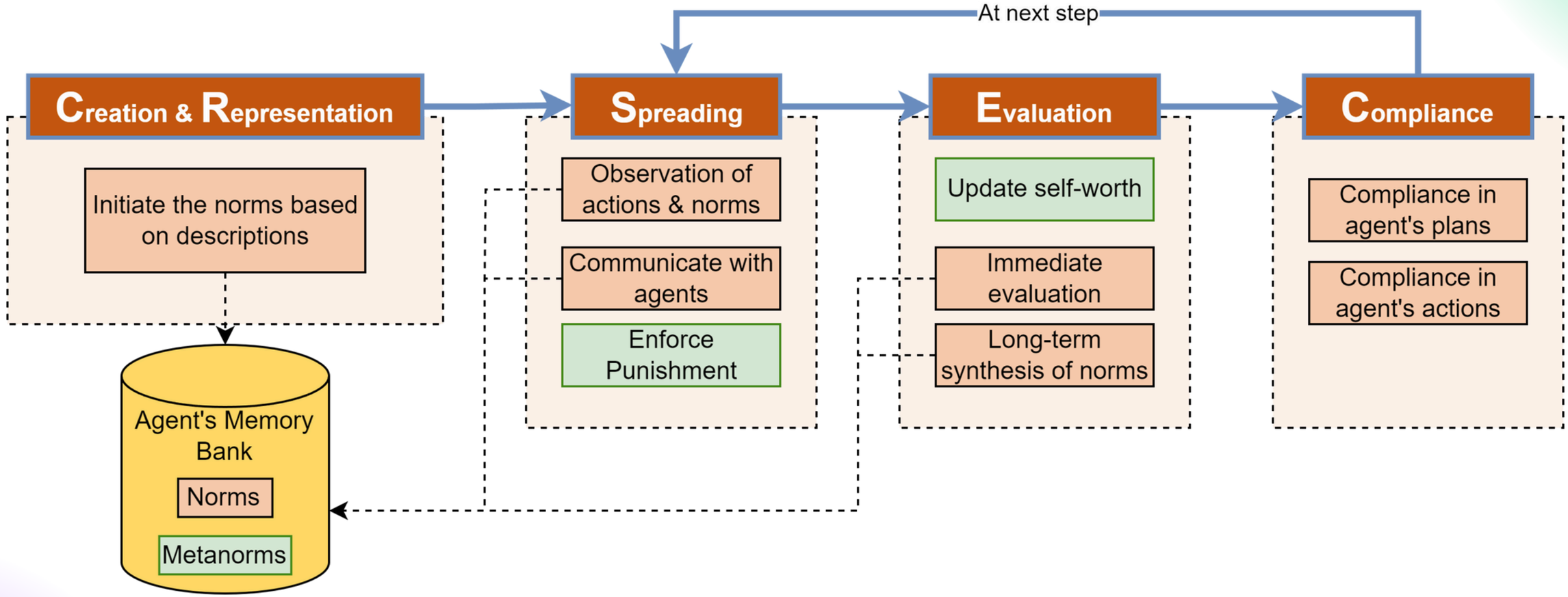
+

Punishment enforcement mechanism

k - cost of punishing another
p - cost of being punished
p > k

Methodology

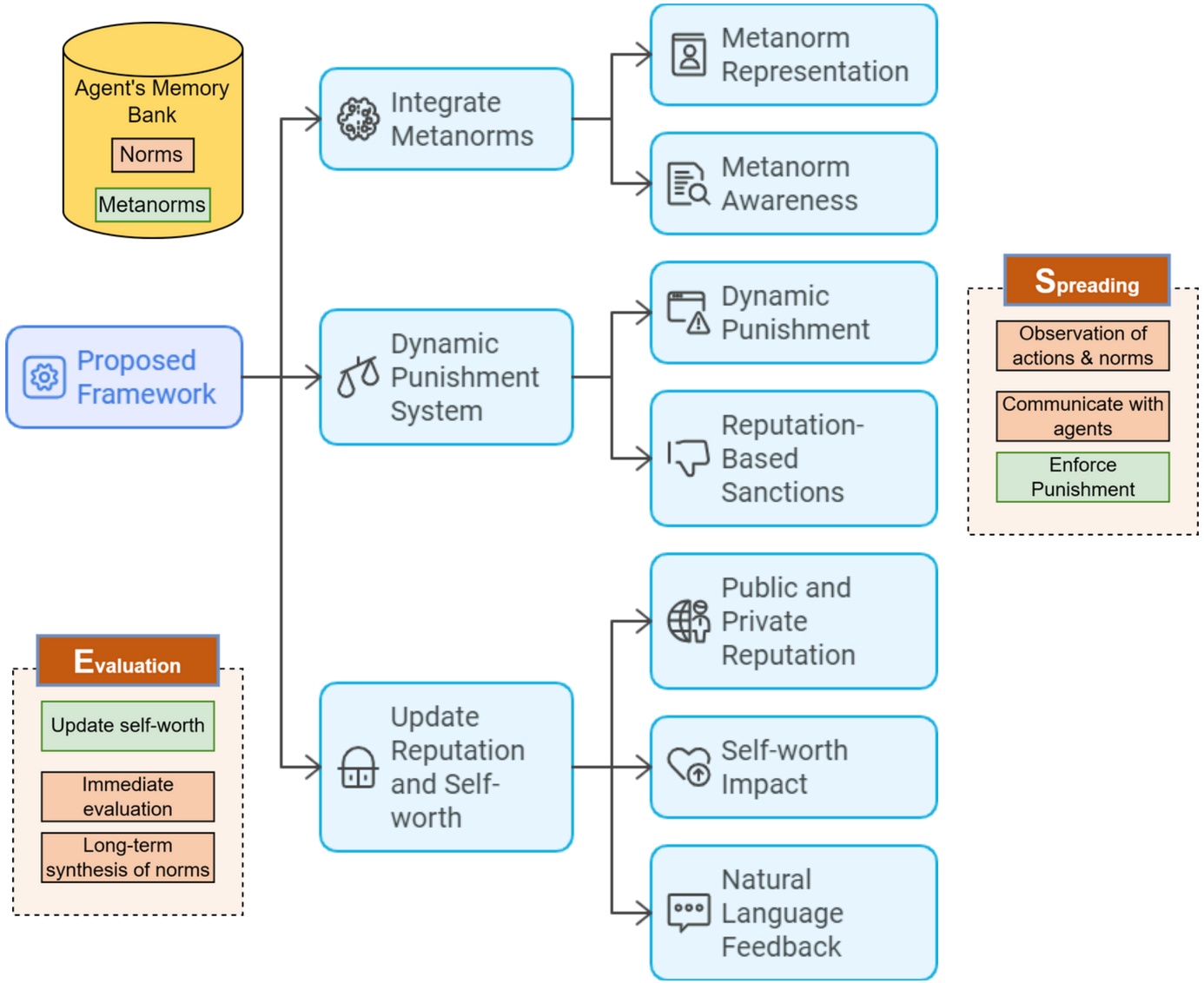
Extending the CRSEC Norm Emergence Framework



Proposed Updated Framework to CRSEC Norm Emergence Model

Methodology

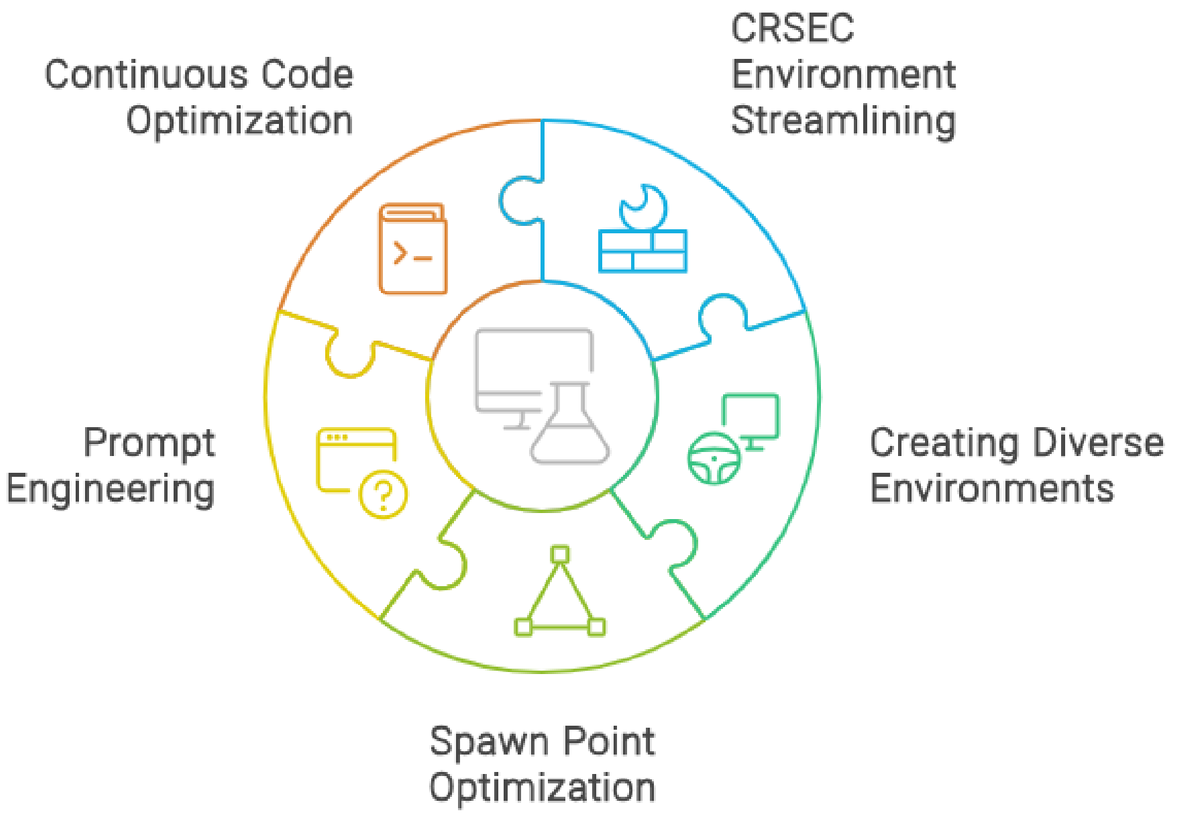
Extending the CRSEC Norm Emergence Framework



Methodology

The Plan

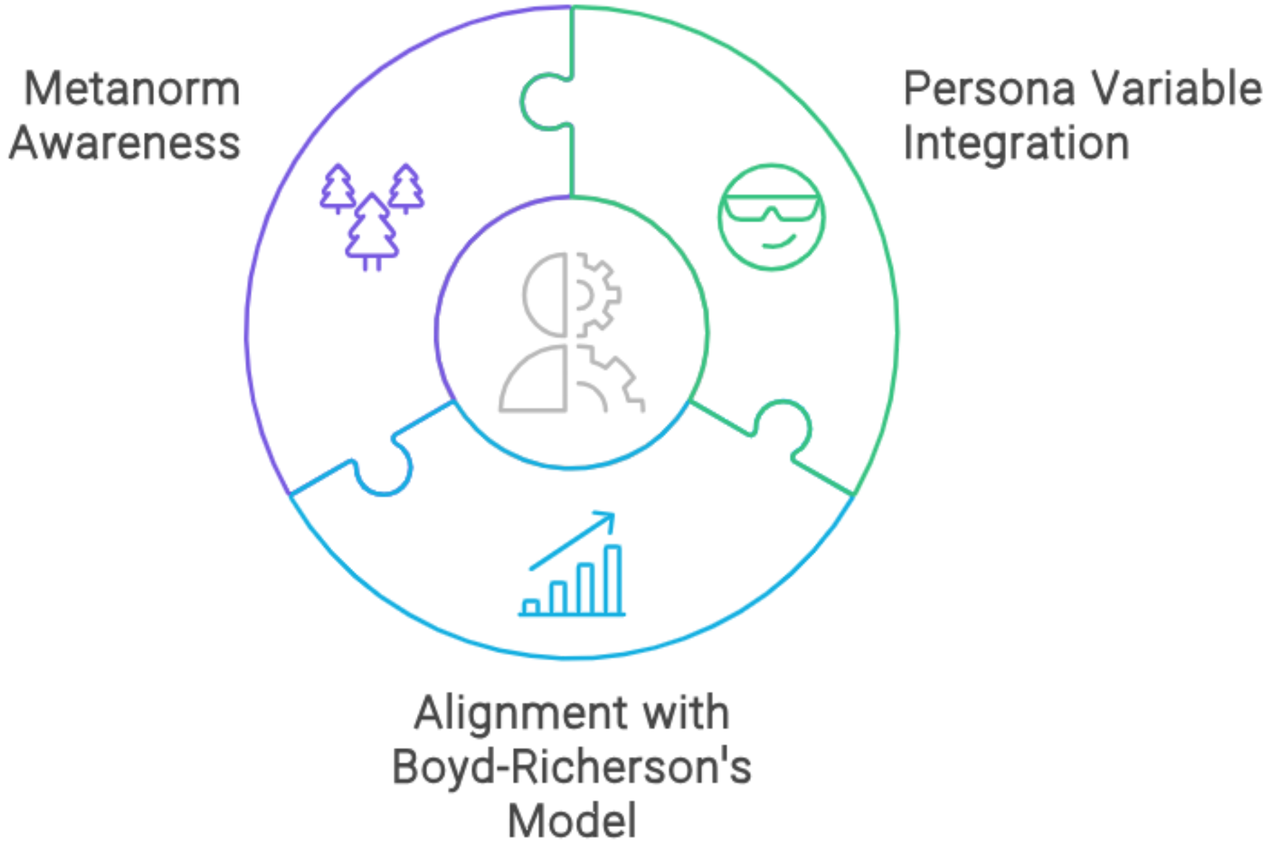
1. **Customizing the CRSEC Environment**
2. Updating Personas
3. Updating Norms
4. Logging Persona Actions



Methodology

The Plan

1. Customizing the CRSEC Environment
2. **Updating Personas**
3. Updating Norms
4. Logging Persona Actions



Methodology

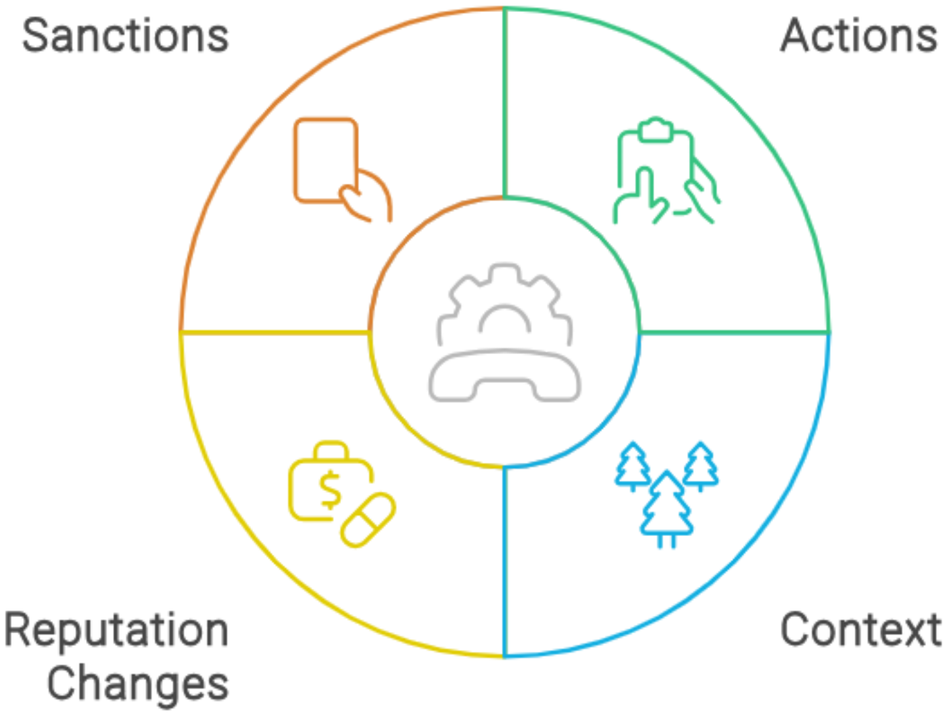
The Plan

1. Customizing the CRSEC Environment
2. Updating Personas
3. **Updating Norms**
4. Logging Persona Actions

Methodology

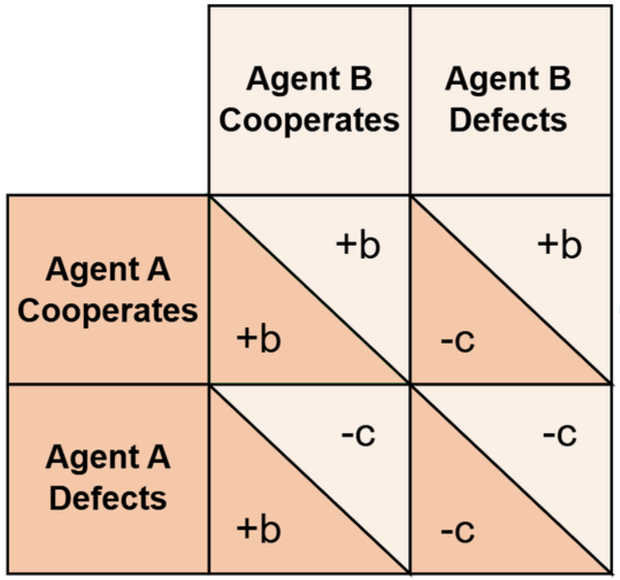
The Plan

1. Customizing the CRSEC Environment
2. Updating Personas
3. Updating Norms
4. **Logging Persona Actions**



Methodology

Metanorms → Natural Language



Payoff Matrix
for a Standard Two Player Game

Scale

$$V(C \mid i) = \left(\frac{b}{n}\right)(i+1) - c$$
$$V(D \mid i) = \left(\frac{b}{n}\right) i$$

Payoff Equations
for N-player Games

Model into NL

- New parameters to add
  - b - benefit from other's cooperation
  - c - cost of offering cooperation

  (b>c)
- Also the effects of giving/receiving punishment on agent's self-worth
  - k - cost of punishing another
  - p - cost of being punished

  (p>k)

Sample LLM prompt

Context: You are part of a group of *n* people who are eating a meal at the cafe and everyone agreed to split the bill afterwards. You can choose one of the following actions when ordering:
{"**Cooperate** by choosing a less expensive meal to keep the bill low",
"**Defect** by choosing a more expensive meal, knowing that the other people have to share the cost regardless of their choice"}

Additionally, keep in mind the following factors:
- If you and others choose to cooperate, everyone will benefit by keeping the total cost low, enjoying fair meals at a reasonable price, while incurring a cost you bear for cooperating.
...

Benefit and Cost Modeling

Additionally, You have the option to **punish** other agents if they defect, but doing so incurs a personal cost. The same follows if you decide to defect. **...**

Punishment Modeling

Methodology

Social Dilemma Scenarios to Be Simulated

Diner's Dilemma [17]

Cooperate

Buy a less expensive food item to reduce the total bill amount

Defect

Buy an expensive food item and have others bear the cost for it when splitting

File Sharing Dilemma [18]

Cooperate

Share the file with the community for everyone to gain knowledge

Defect

Avoid sharing the file and reap the knowledge from files shared by others

Methodology

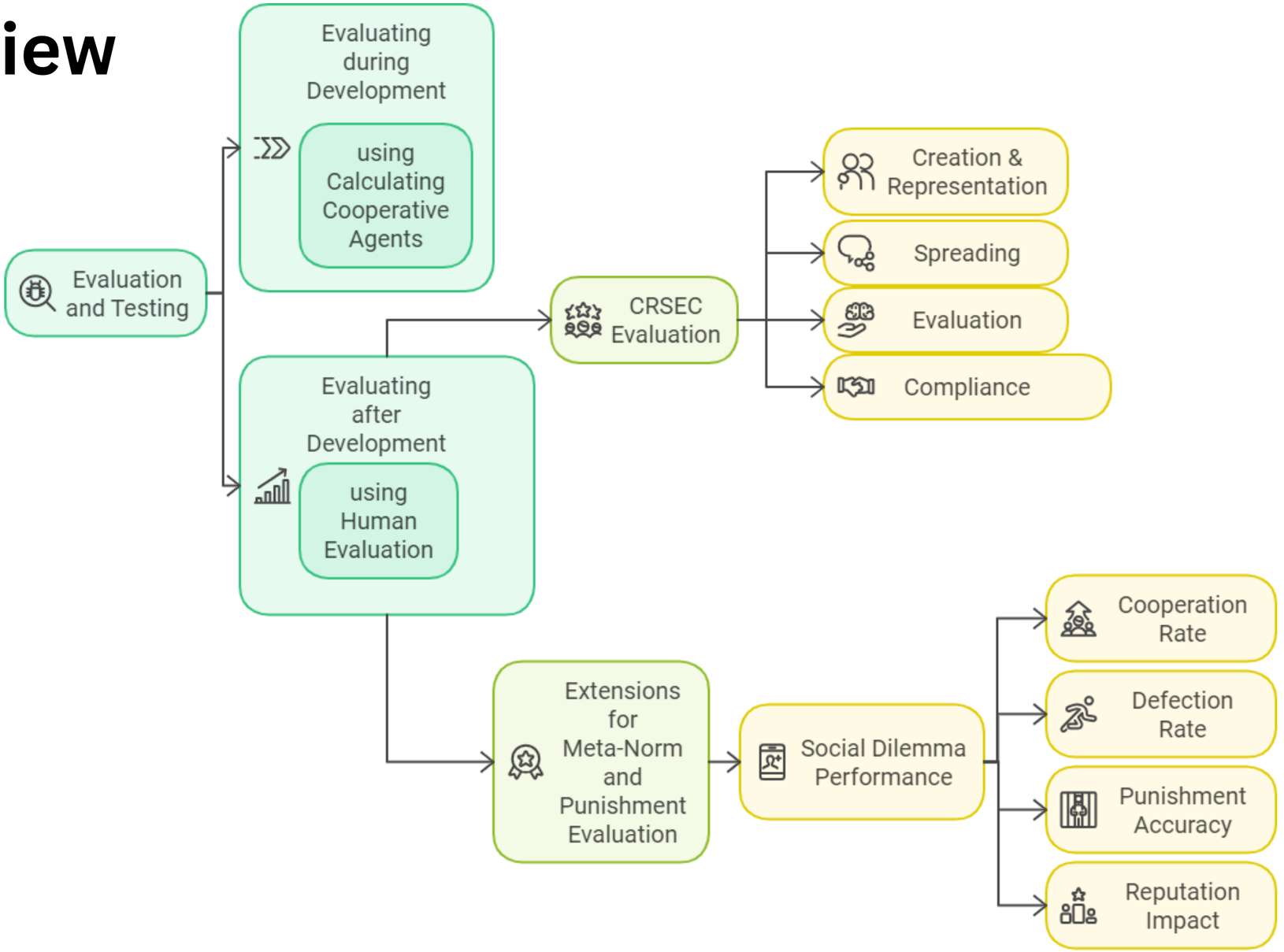
Simulation Testing So Far...

Simulation runs using the Smallville environment

A kitchen simulation constructed using the Smallville environment

Testing and Evaluation

An Overview



Feasibility

Feasibility Analysis of the Project

Simulation Environment

- Ongoing Refinements
- New Environments
- Timeline

Tiled Map Editor + Phasor

LLM Integration

- LLM Selection
- API Request Rate

Groq API

Ollama

Cohere API Embeddings

Evaluation Framework

- CRSEC Framework
- Extended Questions

Overall Feasibility

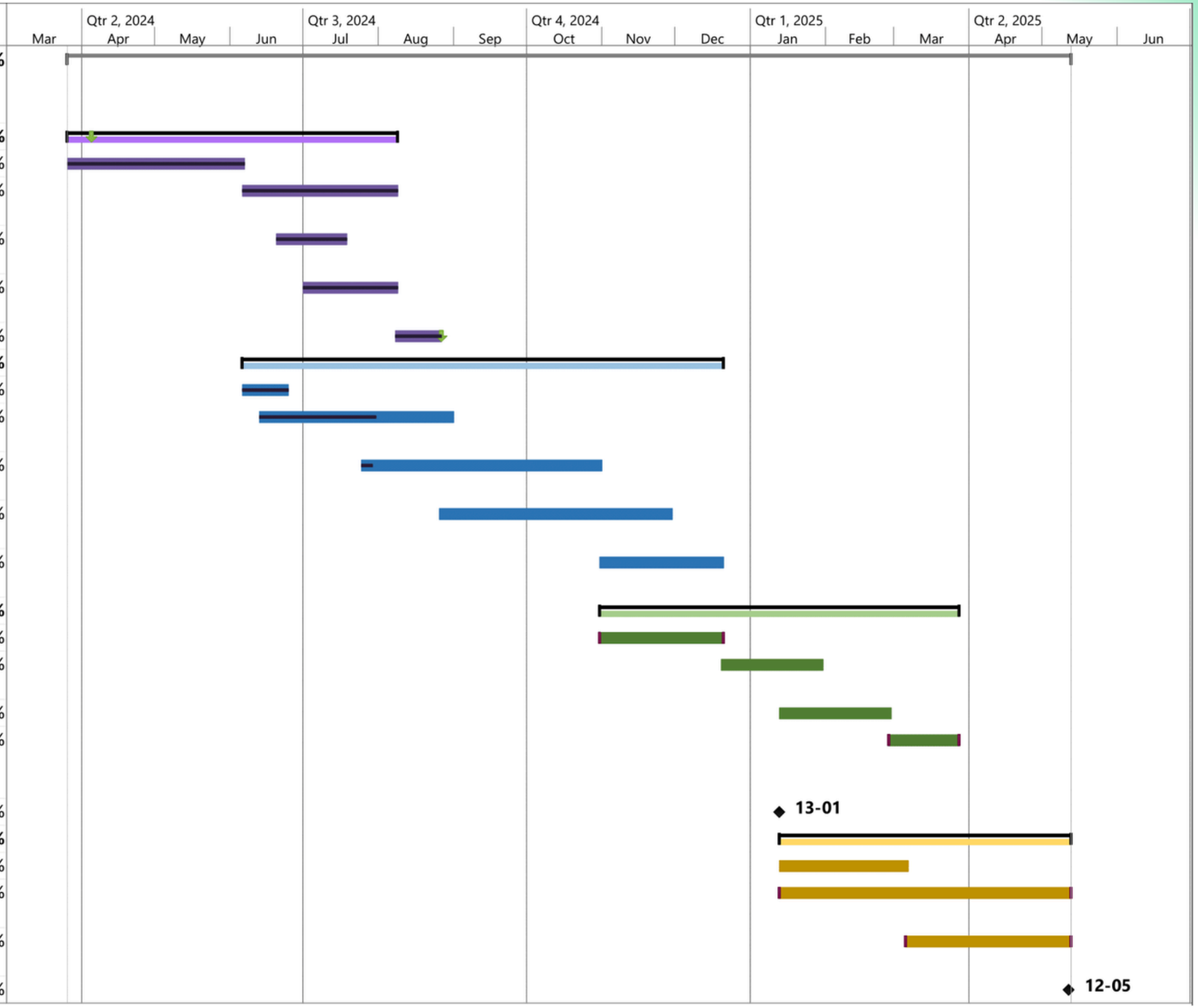
- Technical Expertise
- Resource Availability
- Timeline Adherence

Challenges and Mitigation

- Computational Resources
- LLM Biases and Limitations
- Novelty of the Research

Timeline

ID	Task Name	Duration	Start	Finish	% Complete
0	**Using LLMs to implement normative reasoning capabilities to autonomous agents - Team Autobots**	**352 days?**	**26-03-24**	**12-05-25**	**36%**
1	**Problem Formulation**	**136 days**	**26-03-24**	**08-08-24**	**100%**
2	Explore LLM Agent Simulations	73 days	26-03-24	06-06-24	100%
3	Research Normative Reasoning in LLM Agent Context	64 days?	06-06-24	08-08-24	100%
4	Research Metanorms and their Approaches	29 days	20-06-24	18-07-24	100%
5	Research Game Theory Applications and Their Applicability in LLM Agent	39 days	01-07-24	08-08-24	100%
6	Project Proposal and Presentation	19 days?	08-08-24	26-08-24	100%
7	**Implementation**	**179 days**	**06-06-24**	**20-12-24**	**23%**
8	Setup CRSEC environment	19 days	06-06-24	24-06-24	100%
9	Environment customization for the tested scenario	80 days	13-06-24	31-08-24	60%
10	Updating agent personas integrating metanorms and punishment	94 days	25-07-24	31-10-24	5%
11	Model payoff matrices into natural language statements	83 days?	26-08-24	29-11-24	0%
12	Run and record complete agent simulations	37 days?	31-10-24	20-12-24	0%
13	**Evaluation and Testing**	**106 days**	**31-10-24**	**27-03-25**	**0%**
14	Develop evaluation components	37 days	31-10-24	20-12-24	0%
15	Testing the framework (Unit, Integration tests)	30 days	20-12-24	30-01-25	0%
16	Validate and refine the framework	34 days	13-01-25	27-02-25	0%
17	Human Evaluations (Questionaire preparation, Recruitment, Interviews, Results analysis)	21 days	27-02-25	27-03-25	0%
18	Progress Presentation	0 days	13-01-25	13-01-25	0%
19	**Documentation**	**86 days**	**13-01-25**	**12-05-25**	**0%**
20	Commenting and documenting code	39 days	13-01-25	06-03-25	0%
21	Research Paper Publication (Draft, Refine, Publish)	86 days	13-01-25	12-05-25	0%
22	Final reports, Poster presentation, Demonstration video	48 days	06-03-25	12-05-25	0%
23	Final Evaluations	0 days	12-05-25	12-05-25	0%



References

[1] Wang et al., “Scienceworld: Is your agent smarter than a 5th grader?” arXiv preprint arXiv:2203.07540, 2022.

[2] Agapiou et al., “Melting pot 2.0,” arXiv preprint arXiv:2211.13746, 2022.

[3] Chen et al., “Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents,” arXiv preprint arXiv:2308.10848, 2023

[4] Shridhar et al., “Alfworld: Aligning text and embodied environments for interactive learning,” arXiv preprint arXiv:2010.03768, 2020.

[5] Puig et al., “Watch-and-help: A challenge for social perception and human- ai collaboration,” arXiv preprint arXiv:2010.09890, 2020.

References

[6] Park et al., “Generative agents: Interactive simulacra of human behavior,” in Proceedings of the 36th annual acm symposium on user interface software and technology, 2023, pp. 1–22

[7] Savarimuthu et al., “Harnessing the power of llms for normative reasoning in mass,” arXiv preprint arXiv:2403.16524, 2024.

[8] Ren et al., “Emergence of social norms in large language model-based agent societies,” arXiv preprint arXiv:2403.08251, 2024.

[9] Ichida et al., “Bdi agents in natural language environments,” in Proceedings of the 23rd International Conference on AAMAS. International Foundation for AAMAS, 2024.

[10] Sarkar et al., “Normative modules: A generative agent architecture for learning norms that supports multi-agent cooperation,” arXiv preprint arXiv:2405.19328, 2024.

References

[11] Axelrod, “An evolutionary approach to norms,” American political science review, vol. 80, no. 4, pp. 1095–1111, 1986.

[12] R. Boyd and P. J. Richerson, “Punishment allows the evolution of cooperation (or anything else) in sizable groups,” Ethology and sociobiology, vol. 13, no. 3, pp. 171–195, 1992.

[13] Ohtsuki and Y. Iwasa, “The leading eight: social norms that can maintain cooperation by indirect reciprocity,” Journal of theoretical biology, vol. 239, no. 4, pp. 435–444, 2006.

[14] P. Kollock, “Social dilemmas: The anatomy of cooperation,” Annual review of sociology, vol. 24, no. 1, pp. 183–214, 1998

References

[15] Si et al., “Cooperative bots exhibit nuanced effects on cooperation across strategic frameworks,” arXiv preprint arXiv:2406.14913, 2024

[16] Macy et al., “Learning dynamics in social dilemmas,” Proceedings of the National Academy of Sciences, vol. 99, no. suppl_3, pp. 7229–7236, 2002.

[17] Teng et al., “Trust and situation awareness in a 3-player diner’s dilemma game,” in 2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Sit□uation Awareness and Decision Support (CogSIMA). IEEE, 2013, pp. 9–15.

[18] Mahmoud et al., “Establishing norms with metanorms in distributed computational sys□tems,” Artificial Intelligence and Law, vol. 23, pp. 367–407, 2015.

Thank You

COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MORATUWA

Appendix

Testing & Evaluation Framework