AI SAFETY CAMP 2026

Compute Permit Markets under Imperfect Monitoring

Week 2: Progress Review & Scope Decision

Co-Mentors: **Joel Christoph** & **Jonas Kgomo**

22 January 2026

This Week's Agenda

Review

- New documents submitted
- Key insights from Week 1 work
- Open questions surfaced

Decide

- Scope: Theory-first vs. simulator-first?
- How to proceed with Apart sprint (Jan 30–Feb 1)

Documents This Week

- `Constraints_faced_by_regulators`
- `Week_1_Notes_Audit_Parameters`
- `JTuffy_V0_Diagram`

Plan

- Tasks for next week
- Share outputs in Google Drive folder

Sue: Enforcement Constraints

File: `Constraints_faced_by_regulators.docx`

EU ETS Constraints (Transferable)

1. Self-monitoring and self-reporting
2. Measurement uncertainty
3. Limited audit capacity
4. Decentralized enforcement
5. Boundary definitions limit observability

AI-Specific Constraints (Harder)

1. Extreme information asymmetry
2. Distributed transnational infrastructure
3. No standardized risk metrics

Sources: EU MRR 2018/2066, AVR 2018/2067

Takeaway: AI compute faces EU ETS constraints plus higher asymmetry and boundary fluidity.

Emlyn: Audit Parameters

File: `Week_1_Notes_Audit_Parameters.pdf`

Core Deterrence Condition

$$pB \geq q$$

where p = detection prob., B = penalty, q = gain from non-compliance.

With Backchecks

$$p_{\text{eff}} = p_s + (1 - p_s) \cdot p_b$$

p_s = audit success, p_b = backcheck prob.

Statistical Detection Model

- Auditor sees noisy signal: $\hat{T} = T + \varepsilon$
- Detection if: $|R - \hat{T}| > \tau$
- Study false negatives / true positives

Test Cases Identified

- Bunching below thresholds
- General underreporting
- Failed backchecks

High priority: Backchecks as simulator parameter. Reputation dynamics as cost factor.

Josh: Simulation Architecture

File: `JTuffy_V0_Diagram.svg`

High-Lift Components

- **Actor behavior:** Incentives, optimal strategy computation—possibly expected value, but complex
- **Permit market:** New action space, price resolution each turn, strategies based on $t - 1$ prices

Lower-Lift Components

- Banking
- Price collars
- Thresholds
- Grandfathering

Can be added onto working MVP.

Josh's view: "A complete simulation will be substantial work." Core components may not be tractable.

Decision Point: Theory-First vs. Simulator-First

Joel's Proposal (Jan 20): Narrow to *theory of compute permit markets under imperfect monitoring*.

Theory-First

- Decision-relevant conclusions about market and enforcement designs
- What works when monitoring is noisy, audit capacity limited, actors strategic?
- Toy examples to illustrate, not a large simulator
- Tighter coordination, clearer division of labor

Simulator-First

- Build working MVP, see if core is tractable
- Risk: substantial work, uncertain payoff
- Emlyn: "Either way we need theory on paper first"
- Could revisit in 2 weeks based on progress

Open Questions

Modeling Questions

- How do we model actor strategy computation? Expected value, or more complex?
- How do we handle market price resolution each turn?
- What is the right noise model for auditor observations?
- How do we parameterize firm-specific detection probabilities?

Scope & Feasibility Questions

- Is the simulator tractable in 12 weeks?
- Should we commit to theory-first now, or wait 2 weeks?
- What minimal experiment validates the theory?
- How do we use the Apart sprint (Jan 30–Feb 1)?

Decision needed today: Do we commit to theory-first, or run a 2-week simulator feasibility test?

Upcoming: Apart Research Sprint (Jan 30 – Feb 1)

What It Is

- 3-day intensive research sprint
- Opportunity for focused, uninterrupted work
- Good forcing function for MVP

Who Can Join?

- Open to team members with availability
- Coordinate via Slack by Jan 27

Possible Sprint Goals

- **If theory-first:** Draft core theoretical framework, define key propositions
- **If simulator-first:** Build minimal working prototype, test actor behavior module
- **Either way:** Produce tangible artifact by Feb 1

Action: Decide sprint participation and goals by Tuesday Jan 27.

Tasks for Next Week (by Thursday 29 January)

Everyone: Confirm sprint availability. Upload outputs to Google Drive.

Person	Task
Sue	Expand constraints doc: prioritize which are most binding. Suggest 2–3 for theory.
Emlyn	Formalize deterrence model in short note. Define simulator parameters.
Patrick	Draft simulator architecture doc based on Josh's diagram. Identify MVP modules.
Josh	Define 2–3 concrete scenarios with parameter values for MVP testing.
Adebayo	Literature summary: 3–5 key papers on permit markets under imperfect monitoring.
Pilar	Stakeholder mapping: who uses our memo? What questions do they have?

Sprint prep: Finish Week 2 tasks by Jan 29 so sprint time is productive.

Summary and Next Steps

This Week's Progress

- Sue: EU ETS constraints mapped
- Emlyn: Audit detection formalized
- Josh: Architecture diagrammed

Decisions Today

- Scope: theory-first or 2-week test?
- Apart sprint goals and participation

Next Meeting

Thursday 29 January, 13:00 UTC

Pre-sprint check-in. Review Week 2 outputs.

Reminder

Please upload all outputs to the shared Google Drive folder so everyone can review.

Questions?

Joel Christoph & Jonas Kgomo

jonachro@gmail.com | jonas@equiano.institute