# AISC Week 2 - The Deterrence Model and Simulator Parameters

Emlyn Graham

## Model summary

*Note: I have included caveats and requests for feedback throughout the document.*

**Compliance condition:** Firms will comply if the expected value of the compliance is greater than that of non-compliance

$$p \times B \geq g$$

where:

- $p$ = detection probability
- $B$ = penalty for being caught
- $g$ = gain for non-compliance

## Model breakdown

I have nested these by factor, and expect for the simulator we could start with the most simple and then split out the sub-factors from there.

### Detection probability breakdown

Detection is a function of the regulator (audit capacity, backchecks, other monitoring) and the firm (whistleblowing rate). Regulator capacity may be tied to market revenue in a later stage.

- $p = p_a + p_{other}$
  - $p_a \rightarrow p_{a(i)}$ = base audit rate (for example, may audit 20% of firms each period). Should assume constant to begin with.

        * $p_{a(i)} \rightarrow c(i)p_a$ = firm-specific audit rate, taken from the base rate and scaled by some firm coefficient $c(i)$ for firm $i$ representing other decision-making factors (maybe oversight is dependent on firm size, track record, etc.)
        * $p_a \rightarrow (1-\epsilon_{II})p_a + \epsilon_{II}p_b$ = add type-II error $\epsilon_{II}$ (audit fails to catch misreporting) and backcheck rate $p_b$ which will catch some of these.
    - $p_{other} = p_w + p_m$ = other detection factors
        * $p_w = p_w(i)$ = the whistleblower detection rate, likely relative to a firm's size and AI safety culture
        * $p_m$ = global monitoring, including hardware metering, electricity metering. *May be relative to the size/advanced level of the firm.*

**Penalty breakdown**

Penalty is a function of the regulator (financial or legal penalty) and the firm (sensitivity to reputational damage).

- $B = B_r + R$
    - $B_r = B_f + B_l$ = regulatory penalty (fixed or variable, financial and/or other legal) (higher regulatory penalty, more likely to comply) *It may be easier to not break these up if we don't want to separately model these sub-factors.*
        * $B_f$ = financial penalty (fines)
        * $B_l$ = legal penalty (legal costs and criminal liability)
    - $R = R_e + R_x$ = reputation cost if caught (probably firm-specific) (higher reputation cost, more likely to comply) *It may be easier to not break these up if we don't want to separately model these sub-factors.*
        * $R_e$ = economic reputation damage (valuation, lost contracts) (relative to firm size)
        * $R_x$ = existential reputation damage (blacklisting, loss of compute access with chipmakers, shutdown by shareholders) (relative to firm size)

**Gain for cheating breakdown**

Values are a function of the market (pricing) and the firm (underreporting, compute that requires permits).

- $g = \Delta C + V$
    - $\Delta C = C(real) - C(reported)$ = permit cost savings from underreporting
        * $C(real)$ = cost of permits at level of real compute activity (higher cost, less likely to comply)

    - $C(reported)$ = cost of permits at the level of reported activity (which is zero if below threshold required for compute permits) *I would appreciate feedback on this point. It seems to me that this would always lead to zero/minimal reporting (if choosing to cheat) unless we set the detection probability to be a function of* $\Delta C$.
    - $C(x) = C_m(x)$ if above regulatory threshold; $= 0$ if below regulatory threshold. $C_m(x)$ is a function of the market price (including any collar) and banking for individual firms.
  - $V = c_r \times V_b$ = value of model capabilities. *Again I would appreciate feedback here. I suggested modelling as a sum* $\Delta C + V$ *since I see cost savings and value as two separate factors, but in fact the level of underreporting would be dependent on the value seen in the training run.*
    - $V_B$ = baseline value of the model capabilities
    - $c_r$ = racing factor (around $c_r = 1.1$ normally, may be $c_r >> 1$ in a full race to AGI.) *Again not sure whether a scaling factor or a separate added term is most appropriate here.*

## Example parameter values and tests

| Parameter | Suggested range | Source to reference |
|---|---|---|
| $p_a$ | $[0.05, 0.30]$ | Duflo et al. (2013) and EU ETS (as per Sue's analysis, limited capacity) |
| $p_b$ | $[0.10, 0.40]$ | Duflo et al. (2013) |
| $\epsilon_{II}$ | $[0.10, 0.40]$ | *Please provide* (seems this should be high as per Sue's analysis) |

Example tests:

1. Firms at/below threshold always comply; firms with large runs above threshold need higher p to comply.
2. Compliance rate should increase as $p$ and/or $B$ increase.
3. Backcheck rate should significantly reduce misreporting.

## Functioning of the model

*For discussion/feedback.*

I propose that we should first get the model working as defined above in the simplest case, with fixed probabilities, and we should study archetypes of AI firms to see what happens and how we can ensure compliance.

Overall it is simple, tractable, produces interpretable compliance thresholds, and should let us answer the question "what p and B combinations ensure compliance across firm types?"

Issues:

- Won't allow us to properly study spread of misreporting, since firms will either comply or maximise their savings by reporting exactly at the zero cost threshold.
- May be noise in firm decision-making as well, since they cannot fully know the detection probabilities and non-financial penalties.

We should quickly do this and then add the market mechanism with multiple firms and repeated rounds to see system dynamics.

## Extra questions

- The threshold seems to be a key factor we haven't discussed much yet. I assume it can be set at one of the levels proposed in the literature, but we may want it to scale over simulation time. Should check against EU AI Act, US Executive Order
  - Do we want to start with just FLOPs or include intent (model purpose) from the start?
  - Sue identified a question of risk metrics. This is what SaferAI works on, can we pick their brain later in the project?
- How can we include arbitrage in this model?
- Should we consider collusion between firms later?