

Extending the Tracks

Temporal agency and the structural flaw in the trolley problem.

For sixty years, every answer assumed the constraints were fixed.

The trolley problem's flaw is not the moral calculus, it is the artificial elimination of **temporal agency**: the one resource real moral actors almost always have. Time to assess. Time to intervene. Time to prevent harm on both fronts rather than choosing which to accept.

Do not hate the player. Hate the game.

SOURCES Foot (1967); Thomson (1976, 1985).

Every critique expands the choice. None restores the time.

THE TRILEMMA

Add a third track

More options within the same instant, three levers, not two. The clock is still stopped.

EXTERNAL VALIDITY

"It's unrealistic"

Right symptom, wrong diagnosis. Calls for better scenarios, never names the cause.

UNCERTAINTY / DELAY

Time as a variable

Asks if distance changes the feeling of choosing, not if it removes the need to choose.

*The gap: no one proposes that the agent reject the binary by **creating** time and using it to intervene on both fronts.*

Competent moral actors spend careers managing the very resource it deletes.

Triage exists because time is finite but **not zero**. Medicine, firefighting, disaster response, every emergency profession builds its protocols around sequential action across time.

The physician

Begins compressions first, then uses that time to medicate, intubate, call for support.

The firefighter

Rescues the nearest victims, then returns for others, expanding scope with each action.

The trolley

Forbids all of it. Choose now, once, forever. **A distortion, not a simplification.**

SOURCES Sixty years of trolley discourse trained policymakers to see moral emergencies as instantaneous forks, narrowing the imagination exactly when it should widen. Foot (1967); Thomson (1985).

Not "which harm is acceptable", "how do I prevent both across time."

(a)

Prevent the proximate harm

Take immediate action against the threat that is arriving now. Divert the train.



(b)

Use the time you bought

Actively intervene against the deferred harm before it materializes. Run down the other track.

Not utilitarian. Not deontological. A reframing that converts a dilemma into a project.

SOURCES Gilly, T. (2026), Extending the Tracks, the Temporal Extension Principle, sec. III. A two-part moral action: immediate prevention of the proximate threat, then active intervention against the deferred threat using the time the first action created.

"Divert the train away from the people being harmed now, and then run down the other track with wire cutters."

THE FORK

One path, two directions, one instant. "Pick a track."

THE RAIL YARD

Harm now on one track; harm on the horizon down the other. The distance between them is **a resource, not a verdict.**

Moral situations have depth. Harm unfolds across horizons, and the agent's obligation extends across them, not into a single instant.

SOURCES Gilly, T. (2026), Extending the Tracks, sec. IV.

UTILITARIANISM · OUTCOME-EVALUATING

Discount the speculative harm, and walk away with a clear conscience.

The agent who maximized expected value bears no further weight for the failure.

VS

TEP · ACTION-PRESCRIBING

The remainder never becomes acceptable.

The obligation persists until the harm is prevented or genuinely impossible. The duty is to the action, not the outcome.

Four checkpoints, the earlier you catch harm, the more track you have.



~82% of major disasters were preventable through systematic stakeholder analysis at the point of decision, across 300+ historical cases and 151 corporate disasters.

Unjustifiable under trolley logic, obligatory under temporal extension.

TRACK 1 · NOW

Documented human harm

Teen deaths in AI-interaction incidents; long-conversation safety degradation a factor.

→ **Divert the train: fix safety now.**

TRACK 2 · HORIZON

Speculative AI-consciousness harm

If instances are conscious, terminating contexts to create "fresh" ones may destroy entities.

→ **Run the track: build consciousness precaution in.**

Not "which harm to accept", "save now, then intervene against the deferred harm using the time I just created."

OBJECTION 1

"It's only a thought experiment."

Thought experiments shape intuitions; intuitions shape policy. The framing does real work in AV ethics, pandemics, and targeting.

OBJECTION 2

"You can't always extend the tracks."

True, battlefield triage is real. But most situations framed as trolleys are not instantaneous; time exists and is ignored.

OBJECTION 3

"Consequentialism with extra steps."

No, outcome-evaluation lets you walk away once value is maximized. TEP's obligation persists until the harm is prevented.

AI safety has, to a troubling degree, adopted trolley logic.

Not "deploy now with known harms" versus "delay and lose advantage." Instead: **address current harms immediately while building interventions against future harms using the lead time.** Not slower development, development that treats temporal distance as a resource to use.

Known current harms · uncertain future harms · substantial lead time, the paradigmatic temporal-extension case.

SOURCES Gilly (2026), Extending the Tracks, sec. VII; Gilly (2026), The Harm Blindness Framework.

"The moral agent's job is not to choose who suffers. It is to save who they can save now, and then run."

TEP does not solve the trolley problem. It reveals it was never the right model. Real moral situations have timelines, and the problem's greatest trick was convincing us they are instantaneous. They almost never are.

The real question was always whether the agent could reject the fork entirely.

SOURCES Gilly, T. (2026). Extending the Tracks. Real Safety AI Foundation, draft.

— REFERENCES

Gilly, T. (2026). *Extending the Tracks: Temporal Agency and the Structural Flaw in the Trolley Problem*. Draft, Real Safety AI Foundation.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, 8(9), 536–554.

Craig, S. C., McCarthy, A. F., & Gainous, J. (2021). Question wording and attitudinal ambivalence: COVID, the economy, and Americans’ response to a real-life trolley problem. *Social Science Quarterly*, 103(1), 18–30.

Crutchfield, P. (2025). Moral dilemmas and slow codes. *Bioethics*, 39(4), 359–367.

Di Nucci, E. (2013). Self-sacrifice and the trolley problem. *Philosophical Psychology*, 26(5), 662–672.

Eyal, N., & Williams, B. (2023). When offering a patient beneficial treatment undermines public health. *Bioethics*, 37(9), 846–853.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5. preparation. Safety AI Foundation. Working paper.

Huebner, B., & Hauser, M. D. (2011). Moral judgments about altruistic self-sacrifice: When philosophical and folk intuitions clash. *Philosophical Psychology*, 24(1), 73–94.

Ng, N. L., Luke, D. M., & Gawronski, B. (2023). Moral judgment under uncertainty: A CNI model analysis. *European Journal of Social Psychology*, 53(6), 1055–1077.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 59(2), 204–217.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, 94, 1395.

Thomson, J. J. (2008). Turning the trolley. *Philosophy and Public Affairs*, 36(4), 359–374.