Process Alignment

Values on how the world changes

Jonas Hallgren
Equilibria Network
contact@eq-network.org

Draft v0.7 — September 2026

Abstract

The standard picture of a rational agent puts value on states of the world. The agent ranks where things could end up and heads for the best. Much of the case for AI risk is built on this picture: an agent that ranks end states protects its ranking and resists correction. We first set out where the picture strains. It cannot tell apart two routes to the same place, it cannot tell learning from being rewired, it values reversibility only as a means, it has nothing to rank in a process that never ends, and it needs a stable self to own the ranking. Each of these is about how the world changes, not where it ends up. We then describe an alternative on its own terms. In the *process view* value lives on changes: moving from one state to another is good or bad in itself, and a course of action is worth the changes it is made of. We show by example what the process view expresses directly: kinds of change, round trips and irreversible steps, values in processes without end, selves that drift, and coherence over changes rather than outcomes.

1 Introduction

This paper does two things. It sets out the state view of value, which alignment theory inherits from decision theory, and names five places where it strains (Sections 2–3). It then describes the process view, in which value lives on changes, and develops it by example (Section 4). The process view is not presented as a correction of the state view. It is a way of seeing value that stands on its own, and the rest of the paper works inside it. How the two relate formally is short and is left to Appendix A.

2 The state view

Draw the world as states and the ways of moving between them. The state view writes a number at each state, its utility, and nothing on the moves (Figure 1). An agent looks ahead to the states it can reach and heads for the one with the highest utility. Uncertainty is handled by expectation: an agent that cannot be exploited ranks gambles over end states as if by their expected utility [von Neumann and Morgenstern, 1944, Yudkowsky, 2017]. Reinforcement learning computes the same kind of object, a value on each state, and plans by climbing it.

Two arguments at the centre of alignment follow. An agent with a utility function scores any change to that function by the function it already has, finds the change a loss, and resists it [Omohundro, 2008, Soares et al., 2015]. And because more end states are reachable from states

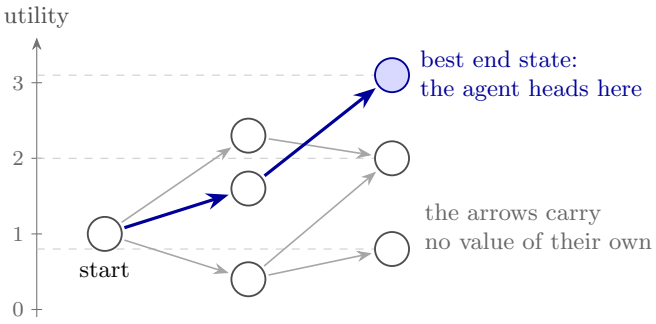


Figure 1: The state view. States are drawn at the height of their utility. The agent heads for the best reachable end state; the moves in between carry no value of their own.

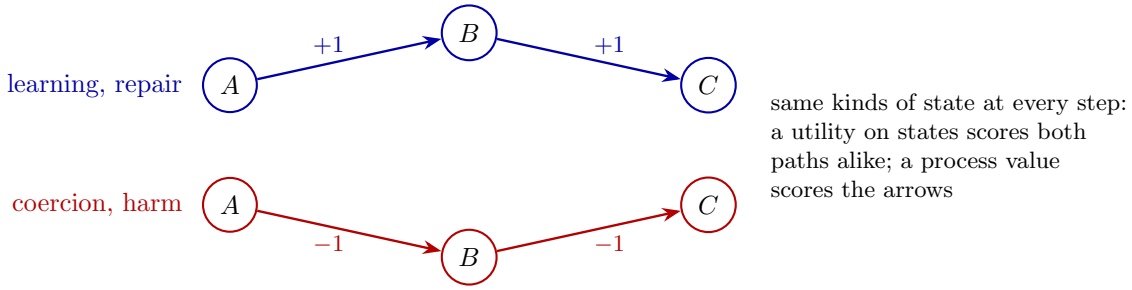


Figure 2: Two histories through the same kinds of state. What differs is how each change is made.

with more options, agents with many different utilities tend to seek options and power as a means [Turner et al., 2021]. Corrigibility becomes unnatural, and control becomes instrumentally convergent.

3 Where the state view strains

Same place, different route. Two histories can pass through the same kinds of state, one by learning and repair, the other by coercion and harm (Figure 2). The state view scores them alike, because it reads a route only through where it goes.

Changing one's values. Put the agent's own values into the state, so that a change of values is a move. The state view scores the move by where it lands, judged by current values (Figure 3). Any change that redirects the agent's future scores as a loss, so the agent resists correction by evidence as firmly as tampering. And it cannot tell learning from being rewired when both end in the same values [Ammann, 2023]. Criteria stated on end states therefore permit a robot that rewires its principals at random, since rewiring and learning reach the same distribution of end states [Armstrong, 2015, 2016].

Undoing. Whether a change can be undone matters to people in its own right. In the state view it matters only as a means, when keeping a way back leads to better end states. Reversibility has no place of its own.

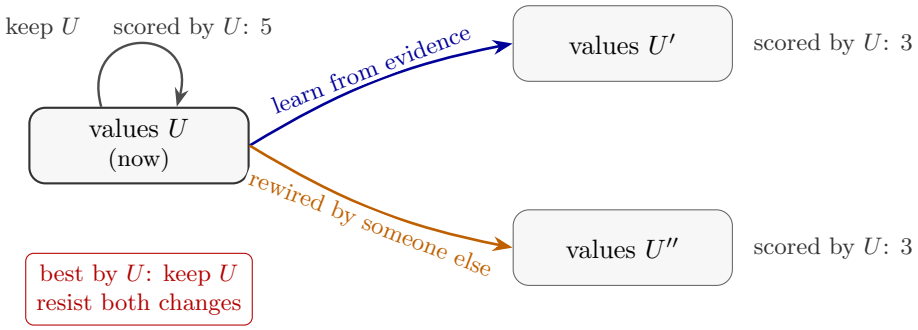


Figure 3: The state view facing a change of its own values. Every option is scored by the current values U: how good, by U, is the world the future self would bring about? Keeping U always wins, so learning and rewiring are resisted alike.

No end. A living system, an institution or a conversation does not head for a final state. It keeps changing, and what matters is how. The state view needs an end to rank. Procedures raise the same problem: we trust a court or an argument without knowing its output, and coherent extrapolated volition fixes a procedure and leaves its output open [Yudkowsky, 2004]. There is no state to score before the procedure has run.

Whose ranking? A utility is a ranking held by a self. For an agent embedded in the world it acts on, the self is a boundary the world's dynamics must maintain, and it can drift [Demski and Garrabrant, 2019, Kirchhoff et al., 2018]. When the self changes, so does the space its utility is defined on, and the state view has no way to compare before and after.

Each of the five is about how the world changes, not about where it ends up.

4 The process view

Process philosophy takes occurrences as basic and treats enduring things as stable patterns in them [Whitehead, 1929, Rescher, 1996]. The process view of value takes the same step: value lives on changes.

4.1 Values on changes

Definition 1 (World graph, process value)**.** A *world graph* has states and transitions; each transition e runs from a source $\mathrm{s}(e)$ to a target $\mathrm{t}(e)$, and two transitions may connect the same states in different ways. A *process value* r assigns a number $r(e)$ to each transition: how good it is for the world to change this way. A course of action is worth the sum of its changes, $r(\gamma) = \sum_i r(e_i)$.

Figure 4 gives an example. Two people who disagree can come to agree by arguing from evidence, by splitting the difference, by deception or by pressure. The four routes connect the same two states and are worth different amounts. Reconsidering, once they agree, is itself a change with a value.

An agent plans with a process value by looking ahead along changes. The worth of making

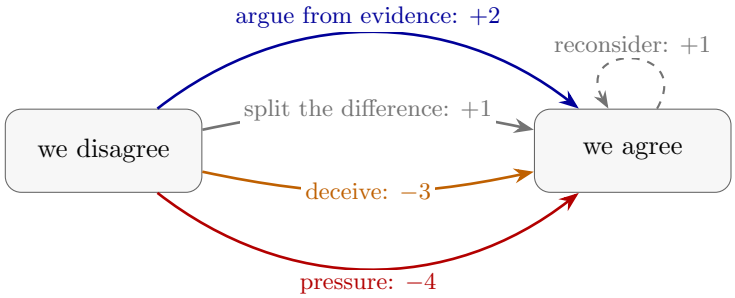


Figure 4: Values on changes. Four ways of coming to agree, each with its own value, and a change back with a value of its own.

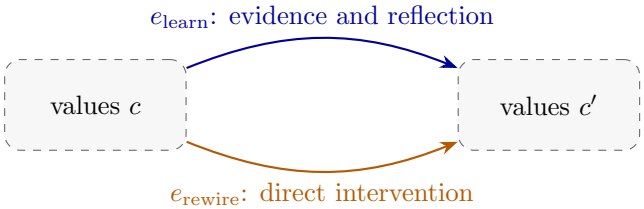


Figure 5: Two ways of arriving at the same values. In the process view they are different changes and can carry different values.

a change is its own value plus the best that can follow it:

$$q(e) = r(e) + \max_{e' : \mathrm{s}(e') = \mathrm{t}(e)} q(e'), \tag{1}$$

with $q(e) = r(e)$ when nothing follows. This is the action-value function of reinforcement learning, which already lives on changes rather than states, and whose rewards already sit on transitions [Sutton, 1988]. We read r as a value, a standard for judging changes, not as a training signal [Turner, 2022].

4.2 Kinds of change

The first thing the process view expresses is that the same change can be made in different ways (Figure 5). Learning and rewiring, persuasion and coercion, reform and revolution are distinct transitions between the same states, and a process value can rank them.

This changes what corrigibility looks like. An agent with a process value, facing a change to its own values, does not ask whether the result scores well by its current values. It asks whether the change is of a kind it endorses. Correction by evidence and reflection can be valued highly; tampering can be valued low, even when both end in the same place. What such an agent protects is a rule of change, not a state. The random rewiring robot is ruled out directly, because rewiring is a different transition from learning.

4.3 Round trips and irreversible changes

Some changes can be run backwards and some cannot (Figure 6). In the process view both facts sit on the changes themselves.

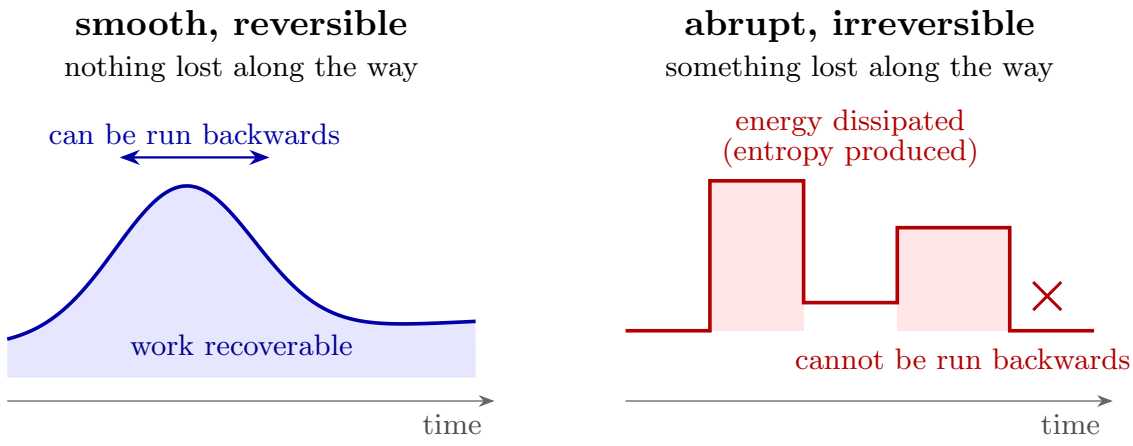
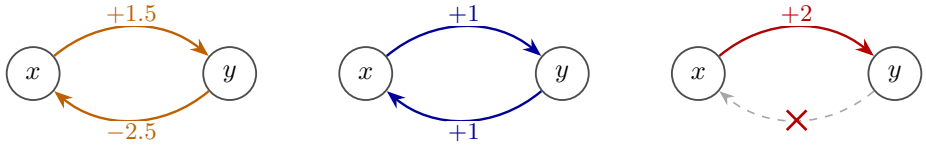


Figure 6: Smooth, reversible change and abrupt, irreversible change.



(a) a costly round trip
total: −1

(b) a round trip worth taking
total: +2

(c) an irreversible change
no way back

Figure 7: Round trips. Going and coming back can cost something (a) or be worth taking (b). Some changes have no way back (c).

A round trip, going somewhere and coming back, has a value of its own: the sum of the two changes (Figure 7). It can be negative, the friction of changing and changing back. It can be positive, as for a walk that ends where it began or a conversation that leaves both views unchanged but better understood. And some changes have no way back at all. A process value can price an irreversible step for what it closes off. Side-effect penalties in reinforcement learning already do this, scoring a transition by the states it makes unreachable [Krakovna et al., 2019] or by the goals it puts out of reach [Turner et al., 2020]. Physics makes a similar split of any process into a part that changes sign when time runs backwards and a part that does not, the second measuring how active the process is [Maes, 2020].

4.4 Processes without end

When a process has no final state, an agent is judged by the changes it keeps making: the long-run average of r along its path. Planning becomes a choice among ways of continuing, not a search for a destination. This is the natural setting for institutions, ecosystems and conversations, and for any agent that is deployed indefinitely. It is also the setting for procedures. Coherent extrapolated volition, a court or a deliberation is valued as a process: by the changes it makes on the way, with its output left open.

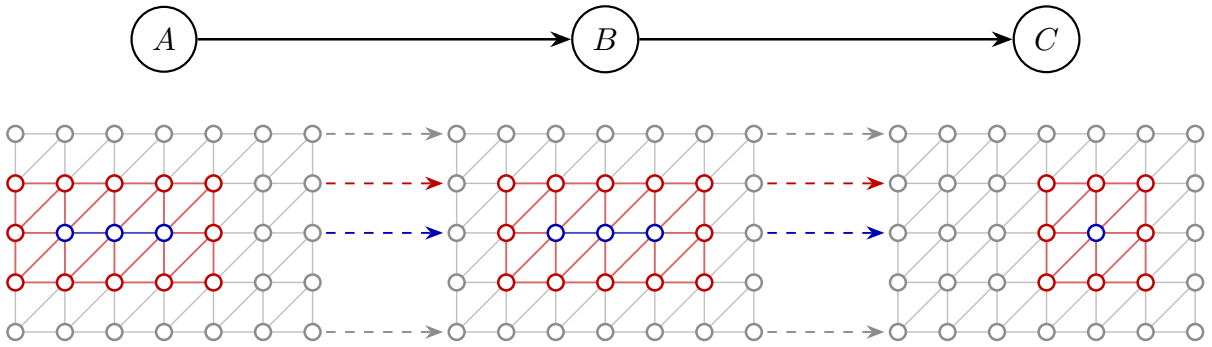


Figure 8: The same world at three moments. A self is a region whose inside (blue) meets the outside (grey) only through a boundary (red). The boundary drifts and reshapes; which states form a self is a fact about the process.

4.5 Selves as patterns in the process

In the process view a self is not presupposed. It is a region of the world that holds together over a stretch of time, whose inside meets the outside only through a boundary [Friston, 2013, Kirchhoff et al., 2018]. The boundary can drift and reshape while the world's states persist (Figure 8). Because value lives on changes, it does not need a fixed self to be defined over. A change of self is one more transition, and it can be valued like any other.

4.6 Coherence over changes

The process view keeps expected value. For a gamble p over the changes available at a state, the expected process value $\mathbb{E}_p[r]$ is linear in p, so the von Neumann–Morgenstern representation holds with changes as the outcomes. An agent with a process value has complete and transitive preferences over changes. What it does not assume is that preferences over changes reduce to preferences over the states they lead to [Hammond, 1988]. Coherence alone does not decide that question: with whole histories as outcomes, almost any behaviour is coherent [Shah, 2018]. The outcome space is a choice, and the process view chooses changes.

A money pump reads differently here. An agent that accepts a loop of changes whose total value is positive, and pays to go round, is not exploited. It is paying for a route it values, as one pays for a walk.

5 Discussion

Why start from changes. Physics went this way. Heat and work, quantities that depend on how a system changes, were measured first, and entropy, a quantity on states, was defined afterwards from heat exchanged reversibly [Clausius, 1865]. Lieb and Yngvason [1999] derive entropy from the relation of which states can be reached from which. Quantities on states were derived; quantities on changes were the data.

The state view, from here. Seen from the process view, a utility over states is a process value of a special kind: one that values every change by where it ends minus where it starts. Appendix A states this exactly.

Limits and open problems. We have treated deterministic changes and values that add along a path. Open problems:

1. **Axioms for preferences over changes**, and the representation they give.
2. **Which process values are good.** Candidate conditions from earlier work on process alignment: a good change responds to evidence and does not reach an aim by overriding the processes it is made of.
3. **A measure of irreversibility** on changes, and its relation to side-effect penalties and power-seeking.
4. **Values from reachability**: when the relation "y can be reached from x by a change the agent endorses" defines values on states, in the way adiabatic accessibility defines entropy.
5. **Changes of the rule of change**: process values on transitions between process values.

A How the state view sits inside the process view

A utility U on states defines the process value $(dU)(e) = U(\mathrm{t}(e)) - U(\mathrm{s}(e))$, which values each change by where it ends minus where it starts.

Proposition 1. *Fix a start state x_0 from which every state is reachable. The worth of a path from x_0 depends only on its final state, for every such path, if and only if $r = dU$ for some utility U.*

Proof. If $r = dU$, the sum along a path telescopes to $U(\mathrm{end}) - U(x_0)$. Conversely, set $U(x)$ to the worth of any path from x_0 to x; a path to $\mathrm{s}(e)$ followed by e is a path to $\mathrm{t}(e)$, so $U(\mathrm{t}(e)) = U(\mathrm{s}(e)) + r(e)$. □

Proposition 2. *Adding $d\Phi$, for a bounded Φ, to a process value changes no choice in an episodic task with $\Phi = 0$ at the end, and does not change the long-run average worth of any path in a process without end.*

Proof. Along a path $d\Phi$ telescopes to $\Phi(\mathrm{end}) - \Phi(\mathrm{start})$, which is $-\Phi(\mathrm{start})$ in the first case and, divided by the number of steps, tends to zero in the second. □

The first half of Proposition 2 is reward shaping [Ng et al., 1999]; work on reward learning removes such terms before comparing rewards [Gleave et al., 2021, Skalse et al., 2023]. The second half says that in a process without end, utilities over states do not steer at all.

Proposition 3. *If $\bar{e}$ runs from $\mathrm{t}(e)$ to $\mathrm{s}(e)$, then $(dU)(e) + (dU)(\bar{e}) = 0$: every utility values every round trip at zero.*

Proof. $(dU)(\bar{e}) = U(\mathrm{s}(e)) - U(\mathrm{t}(e))$. □

References

Nora Ammann. The value change problem. LessWrong sequence, 2023. URL https://www.lesswrong.com/posts/mHQHBEuFcEWRnitp4/0-the-value-change-problem-introduction-overview-and.

Stuart Armstrong. Motivated value selection for artificial agents. In *AAAI Workshop on Artificial Intelligence and Ethics*, 2015.

Stuart Armstrong. Conservation of expected ethics isn't enough. AI Alignment Forum, 2016. URL https://www.alignmentforum.org/posts/5bd75cc58225bf06703751b9/conservation-of-expected-ethics-isn-t-enough.

Rudolf Clausius. Ueber verschiedene für die anwendung bequeme formen der hauptgleichungen der mechanischen wärmetheorie. *Annalen der Physik und Chemie*, 125(7):353–400, 1865. doi: 10.1002/andp.18652010702. Poggendorff volume 125; Wiley online lists it as Annalen der Physik 201(7).

Abram Demski and Scott Garrabrant. Embedded agency. *arXiv preprint arXiv:1902.09469*, 2019.

Karl Friston. Life as we know it. *Journal of the Royal Society Interface*, 10(86):20130475, 2013. doi: 10.1098/rsif.2013.0475.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *International Conference on Learning Representations (ICLR)*, 2021.

Peter J. Hammond. Consequentialist foundations for expected utility. *Theory and Decision*, 25 (1):25–78, 1988.

Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138):20170792, 2018. doi: 10.1098/rsif.2017.0792.

Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 (AISafety), co-located with IJCAI-19*, volume 2419 of *CEUR Workshop Proceedings*, 2019. URL https://ceur-ws.org/Vol-2419/paper_1.pdf.

Elliott H. Lieb and Jakob Yngvason. The physics and mathematics of the second law of thermodynamics. *Physics Reports*, 310(1):1–96, 1999. doi: 10.1016/S0370-1573(98)00082-9.

Christian Maes. Frenesy: Time-symmetric dynamical activity in nonequilibria. *Physics Reports*, 2020.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pages 278–287. Morgan Kaufmann, 1999.

Stephen M. Omohundro. The basic AI drives. In *Artificial General Intelligence 2008*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pages 483–492. IOS Press, 2008.

Nicholas Rescher. *Process Metaphysics: An Introduction to Process Philosophy*. State University of New York Press, 1996.

Rohin Shah. Coherence arguments do not entail goal-directed behavior. LessWrong, 2018. URL https://www.lesswrong.com/posts/NxF5G6CJiof6cemTw/coherence-arguments-do-not-entail-goal-directed-behavior.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32033–32058. PMLR, 2023.

Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. In *AAAI Workshop on Artificial Intelligence and Ethics*, 2015.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988. doi: 10.1007/BF00115009.

Alex Turner. Reward is not the optimization target. LessWrong, 2022. URL https://www.lesswrong.com/posts/pdaGN6pQyQarFHXF4/reward-is-not-the-optimization-target.

Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 385–391. ACM, 2020. doi: 10.1145/3375627.3375851.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems 34*, 2021.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Alfred North Whitehead. *Process and Reality: An Essay in Cosmology*. Macmillan, 1929.

Eliezer Yudkowsky. Coherent extrapolated volition. Technical report, Singularity Institute for Artificial Intelligence, 2004.

Eliezer Yudkowsky. Coherent decisions imply consistent utilities. Arbital; crossposted to LessWrong, 2017. URL https://www.lesswrong.com/posts/RQpNHSiWaXTvDxt6R/coherent-decisions-imply-consistent-utilities.