

# WHO STEERS WHO STEERS?

## A NOTE ON IDENTIFYING VULNERABLE MORAL PROPENSITIES

Steven Kaas & Steve Rayhawk  
stevenkaas@gmail.com

*There are a variety of processes that steer the future; that is, they move it toward certain states and away from others dynamically, with changing behaviors in response to changing conditions. Our decisions now don't just steer the future directly, but influence what the major steering processes in the future will be. Certain dangers attend such a project. Often the replacement of part of an ecosystem or a society with an engineered substitute, designed on the basis of a partial understanding, meets with severe drawbacks from unrecognized missing functionality that was demolished or displaced. In the same way, a project to take into hand the steering of the future, to fulfill its potential according to some moral vision, risks demolishing or displacing some unrecognized steering processes that generated and preserved the correctness of the moral sense behind the vision. While this risk cannot be avoided entirely, it can be mitigated by developing better tools for identifying or avoiding interference with unrecognized steering processes. In light of modern physical ontology, and in light of the abstractness of some of the plausible processes (such as the relative market success of firms, or historical evolutionary selection), we suggest that such tools should be rooted in a very conservatively general theoretical framework, based on finding factorizations of the world's state space into potential "steering" and "steered" state subspaces, with partially coupled dynamics.*

*Quis custodiet ipsos custodes? (Who will guard the guardians themselves?)*

---

Juvenal, Roman poet, on the problem of engineering stably moral governance

Yo dawg, I heard you like cars, so we put a car in your car, so you can drive while you drive.

---

Xzibit, television presenter, on upgrading a car to contain a racing simulation

Just as our evolutionary history has selected for brains that select courses of action expected to achieve goals previously associated with reproductive success, and just as the authors of constitutions have written laws designed to cause good laws to be written, when we attempt to think about "steering the future", we engage in a project to "drive while we drive": to strategically influence the main factors that will strategically influence the future.

The idea of debating how the future should be steered highlights an inevitable aspect of futurist advocacy that is obscured when one simply debates particular courses of action. First, the future will—just like the present—contain processes that react to the state of the world in ways directing it toward some particular outcomes and away from others, not by reading from a pre-defined script of strategies, but by reacting to observations and new ideas; we'll use this as a rough and provisional definition of "steering processes". But second, the nature and existence of these future processes can depend on present decisions.

"Steering who steers", then, refers to the case where there exists or could exist some entity in the world strategically trying to direct the world to some particular state, and we're acting strategically to determine whether or not this entity exists, or whether it gains or loses influence; or whether or not another entity with similar abilities but different goals will stand in its place.

The project of deciding what kinds of steering processes to allow in the future imposes a certain responsibility. An analogy might be helpful here. Researchers in the field of environmental science have introduced the concept of "ecosystem services" to refer to useful functions that existing Earth systems are performing for us, but that could be disrupted or destroyed. In the 1990s, there was a famous attempt to duplicate such services based on an *a priori* scientific understanding, named Biosphere II. This project ran into severe glitches: among other problems, the project's original engineering goal of a completely closed ecosystem had to be abandoned to prevent oxygen levels from dropping too much (Severinghaus et al., 1994). Similarly, one can find many examples of totalitarian regimes of various scales that tried to extinguish existing processes in their societies on the basis of ideology or central planning, and subsequently generated catastrophic consequences on various scales when it turned out the pre-existing processes performed functions that the engineered processes did not substitute for. We could imagine something similar applying in the case of humanity's potential moral value, where attempts to decide what steers the future ends up excluding some necessary component.

There are multiple kinds of ways the future can end up differing from the present. Some changes are best interpreted as random moral drift, where the past happens to have different moral features than the present in much the same way that different cultures happen to have different moral features. Other changes are best interpreted as genuine moral progress, arising from, e.g., an objectively improved understanding of moral questions.

It is difficult for a naive observer to distinguish between these two scenarios. Just as a culture might be biased to think cultures close to it are more moral simply because they're more similar, we may be biased to believe we are seeing moral progress merely because, over time, the world has been becoming more similar to the present. On the other hand, on any conception of morality that holds to be at least partly objective, the factors that allow us to get moral questions right (e.g. accumulated factual knowledge, philosophical insight, expanding circles of empathy (Singer, 2011)) can predictably increase over time.

Dynamics that promote a higher degree of morality (or whatever else one holds to make futures valuable) can turn up in places one wouldn't think to look. A fictional but clear-cut example is the action of the ropes by which Odysseus had himself bound to his ship's mast; the ropes weren't directly capable of moral cognition, but their state encoded Odysseus's prior conclusion that it would be better if he heard the Sirens' song but lived than if he was lured to his death. The remaining examples will be more contentious, but this may be inevitable, given that the problem at is-

sue is potential failures of our rational *a priori* calculations about moral relevance. The Enlightenment included methodological changes which affected the practice of both epistemology and moral philosophy; the changes in epistemology appear to have been improvements responsible for the scientific and industrial revolutions, which resulted in increased economic and military power of Enlightenment cultures, resulting in turn in increased relative prominence of moral philosophy which were plausibly improved by the Enlightenment influence as well. In some cases, although the opposite has also happened, market forces acting on different firms have contributed to the increased influence of plausibly more accurate moral ideas through altering the relative success of different business cultures. And there are the evolutionary processes which shaped our moral sense, and the more primitive situation-classifying systems on top of which it is built, in the first place.

Meanwhile, we could imagine accidentally discarding morality altogether, losing our potential of getting it back. Advances in technology may be making this more feasible, a possibility raised for example by C.S. Lewis in "The Abolition of Man" (1943). For a proof of the basic principle, consider the possibility of all-out global nuclear war or similar disasters, wiping out all complex life on Earth, and, *a fortiori*, the structures responsible for human moral progress. A reasonable desideratum for a practice of steering the future is that it avoids such scenarios, or at least bounds their likelihood and degree of occurrence in a small and narrow range.

(Perhaps, as the character Winston Smith argues in George Orwell's Nineteen Eighty-four, permanently obliterating morality is not quite so easy, and the "spirit of man" will never be "overcome", making it impossible to form a stably abusive singleton in Bostrom's (2006) sense. If so, the main tragedy resulting from the obliteration of morality is that other actors eventually have to step in, with all the resulting delays and inefficiencies. However, this seems like an optimistic assumption.)

From this position of radical uncertainty, how would we even begin to identify the part of the universe that "carries morality in it", making it dangerous to "take away the reins" from that part?

Some of the steering processes which it may be necessary to recognize are very abstract. We therefore feel that it will ultimately prove necessary to work in a framework that could formally represent any possible proposed analysis of the world into, e.g., "moral perspective sources" and "fallible nature". When this constraint is related to modern physical ontologies, which typically present the world in terms of states and dynamics connecting states at different times, one finds that the most natural way to formulate the concept of an analysis of the world into parts is as some sort of factorization of the physical state space, so that there is a state subspace for the "morality" and a state subspace for the "nature", with partially coupled dynamics.

In other words, it will be necessary to consider (in principle) all possible states of the observable universe, and all possible factorizations of that state space (under the appropriate formalization); and then try to find rules to identify factorizations containing moral processes which plausibly qualify as attempting to steer the future. One can then see if those rules recover the dynamical sources of morality with which we are intuitively familiar, and then in turn see what what other less-intuitive moral processes or philosophical consequences those rules might imply. This project would be similar to the modern projects of situating the physical sciences in a system of in-principle deductions from simple physical laws, or of constructing a model of the contents of mathematics from basic operations of set theory. The discovered rules

might include aspects of algorithmic probability theory (Li and Vitanyi, 2008), for privileging simple factorizations over complex ones. Adapting them to the quantum regime might involve factorizations such as those used by Tegmark for anthropic probabilities and consciousness (Tegmark, 2014).

Coming down from that level of abstraction, such formalisms may help motivate and specify social norms which could predictably protect the activity of existing processes representing moral preservation or improvement, even without anyone in the society knowing what concrete moral beliefs the processes were acting to reinforce or prevent.

We note one final source of conceptual difficulty. Intuitively, the idea of steering the future implies exclusivity: only one process, possibly a compound process, can be considered to be steering at any given time. If a process which influences the future is being steered by another process, then it is the latter process that is ultimately doing the steering. Any formalism that could answer questions about how to affect the steering of the future must therefore take account of reflexive problems associated with its own implementation, or indeed with the process of its having been advocated and selected over other formalisms. (Similar problems of reflexivity in general-purpose decision making are discussed in Barasz et al. (2014)).

The authors are grateful to Mitchell Porter, John Ku, and Paul Christiano for helpful references and discussion.

## REFERENCES

- Barasz, Mihaly, Paul Christiano, Benja Fallenstein, Marcello Herreshoff, Patrick LaVictoire, and Eliezer Yudkowsky. "Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic." Working Paper. 2014. Web. <<http://arxiv.org/abs/1401.5577>>
- Bostrom, Nick. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5.2 (2006): 48-54. Print.
- Lewis, C.S. *The Abolition of Man*. Oxford: Oxford University Press, 1943. Print.
- Li, Ming, and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. 3rd ed. Berlin: Springer-Verlag, 2008. Print.
- Orwell, George. *Nineteen Eighty-Four*. London: Secker & Warburg, 1949. Print.
- Pinker, Steven. *The Better Angels of our Nature*. London: Penguin Books, 2011. Print.
- Severinghaus, Jeffrey P, Wallace S. Broecker, William F. Dempster, Taber MacCallum, and Martin Wahlen. "Oxygen Loss in Biosphere 2." *Trans. Am. Geophys. Union* 75.3 (1994), 33-37. Print.
- Singer, Peter. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton: Princeton University Press, 2011. Print.
- Tegmark, Max. "Consciousness as a State of Matter." 2014. Web. <<http://arxiv.org/abs/1401.1219.v2>>