From Artificial Morality to NERD: Models, Experiments, & Robust Reflective Equilibrium

Peter Danielson[1]

[1]Centre for Applied Ethics, University of British Columbia
pad@ethics.ubc.ca

Abstract

Artificial ethics deploys the tools of computational science and social science to improve the improve ethics, conceived as pro-social engineering. This paper focuses on three key techniques used in the three stages of the research program of the Norms Evolving in Response to Dilemmas (NERD) research group:

1. Artificial Morality. Technique: Moral functionalism -- principles expressed as parameterized strategies and tested against a simplified game theoretic goal.
2. Evolving Artificial Moral Ecologies. Technique: Genetic programming, agent-based modeling and evolutionary game theory (replicator dynamics).
3. NERD (Norms Evolving in Response to Dilemmas): Computer mediated ethics for real people, problems, and clients. Technique: An experimental platform to test and improve ethical mechanisms.

Artificial Ethics

We take ethics to be a mixture of lore and craft, where the lore aspires to social science and the craft to social engineering. We take the goal of *artificial* ethics to be a pragmatic: to deploy the powerful tools of computational science and social science (broadly taken to include cognitive and evolutionary social science) to drive the science and engineering of ethics.

More concretely, this goal has three stages: 1. Demystify ethics by treating moral problems as functional problems. 2. Refocus on realistic mixed populations instead of utopian monocultures of "ethical" and rational agents. 3. Move from focusing on the goal constructing ethical machines to constructing computationally augmented environments for people to manage the democratic ethics of high technology, including negotiating the possible introduction of ethical machines.

Artificial Morality

In a book and series of papers ((Danielson, 1992; Danielson 1992; Danielson 1995a) the Artificial Morality (AM) project modeled the account of ethics developed by David Gauthier (Gauthier 1977; Gauthier 1984; Gauthier, 1988a; Gauthier, 1988b; Danielson 1991; Gauthier 1991; Danielson 2001a). Briefly, Gauthier's proposal for ethics began with the standard decision and game theoretic account of rational choice and recommended additional axioms constraining rational agents to conditional cooperation in prisoner's dilemmas. It became clear when we modeled Gauthier's proposed strategy as a functioning agent opposed by other types of agent, it was either not rational or not very moral. Worse, the strategy we introduced that does better than Gauthier's proposal, by demanding full reciprocity faces computational limitations. We used a simple diagonal result to show that there is no unique reciprocal strategy (Danielson 1995b; Danielson 2002). This result mirrors, at the level of agents, game theory's folk theorem that warns us that there are infinitely many norms in equilibria, or, in (Boyd and Richerson 1992)'s memorable phrase, "Punishment allows the evolution of cooperation (or anything else) in sizable groups."

Evolutionary Artificial Morality

AM tested tiny populations with equal numbers of agents selected for theoretical reasons and hand crafted in Prolog. Clearly there are several ways to improve this method: allow arbitrary populations of agents generated for other reasons. The second stage of the research project, Evolutionary Artificial Morality (EAM) opened up the generator and test of agents. Genetic programming was used to generate agents (Koza 1992; Danielson 1998a) of unexpected kinds and agent-based simulation explored local interaction effects. Later, replicator dynamics was used to simplify the population models (Skyrms 1996; Danielson 1998b). The main result of this project was to confirm the persistence of mixed populations for most interesting public goods problems.[1] That is, one should expect neither that rational free-riders, or "ethical" altruistic unconditional cooperators to ever form pure populations. Only reciprocal cooperators (of some flavor or another) can stabilize cooperation, and they cannot eliminate either unconditional cooperators or defecting free-riders.

These results have been recently generalized and strengthened by the weight of experimental and field data (Ostrom and Walker 2003; Kurzban and Houser 2005). The persistent mix of human cooperative types, which we might label moral polymorphism, is highly significant for

[1] We avoid "social dilemmas" as it reinforces the binary thinking that overlooked the typical tri-partite mix of strategies for so long.

ethics. Not only is reciprocity underrepresented in the theory of ethics, much time has been spent on the debate between the "ethical point of view" and the moral skeptic, which roughly map onto the two minority positions, unconditional cooperation and free riding. Much work remains to get ethics in touch with its real world base: the majority population of reciprocal cooperators and their puzzling moral emotions.

NERD I Deep Moral Data

Norms Evolving in Response to Dilemmas (NERD) is part of the trend to use data from lab experiments and field studies to inform and constrain theory and simulation. Supported by the Genome Canada, and working with several large genomics labs, the NERD-I survey has managed to collect and mine a rich trove of data on how people make serious ethical decisions about problems and opportunities driven by technological and social change (Ahmad et al. In press; Danielson et al. in press; Ahmad et al. 2005). The NERD-I surveys span human bioethics (genetic testing for β-Thalassaemia) and environmental ethics (salmon genomics and aquaculture).

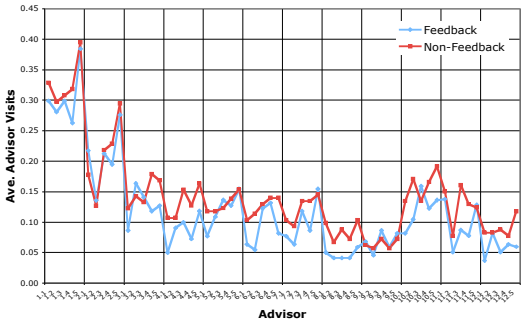

Figure 1 Feedback make little difference

NERD-I is a new form of artificial ethics. Inspired by computer games like Sid Myers' Civilization and Will Wright's Sim* franchise, NERD-I offers constrained choice in a rich context featuring plentiful expert and lay advice. While we do not claim that NERD-I responses are canonical, they have more prima facie normative weight than much of the data gathered by surveys, polls, and focus groups. Moreover, by unifying disparate ethical domains (bioethics and environmental ethics are markedly distinct cultures) NERD-I extends the role of formal methods in the science of ethics. For example, contrary to the common criticism that the framing effect undermines data in ethics, we have show that two different powerful framing stimuli have virtually no effect on responses and advice seeking (see Figure 1).

NERDII Robust Reflective Equilibrium

Nonetheless, NERD-I is very much a social science experiment. Participants are really subjects, given feedback or not, blocked from moving back to earlier answers, and timed at every choice point. Arguably, democratic decision support should have other features.

While we need to know more about our norms, more data cannot solve – indeed deepens – the core problem of ethics. Ethics is fundamentally a coordination problem, of the unstable cooperative kind. Would-be ethical agents need all the help they can affordably get to find partners in beneficial cooperation. This is a communication problem – ethical agents need a common language/protocol for coordination, cooperation, and bargaining, in that order. While they need to be aware of possible complexities and complications, to be told that there are billions of unique personal stories out there is more information than they can use. This was clear from the original Artificial Morality project: reciprocators, the core of any moral solution, need simple recognition schemes and fail miserably under ambiguity. We need ways to coordinate.

Incidentally, this perspective undermines popular criticisms of leading proposals for ethical protocols. To the cost/benefit protocol that effectively drives the most successfully ethics in engineering, epidemiology and public health, and integrated environmental assessment, many critics still fall back on the impossibility of interpersonal comparisons (Sen 2002). But since we are not trying to become ideal impartial spectators but democrats negotiating our policies, we do not face this problem. Put another way, those who cannot tradeoff their own values against those of others are (similar to) sociopaths, not would-be partners in cooperation. People evidently need help to see, not to say calculate, the impact of their casual decisions not to donate blood or solid organs, or to vaccinate their children, but these are well-studied practical problems in risk communication, not symptoms of some deep metaphysical divide between the values of different persons.

We propose to keep philosophy in its place, which is not producing mysterious theoretical problems for ethics, but, we argue, solving the hard practical problems would-be democratic ethical agents face. For example, we need good social networking sites for democratic ethics. Currently it is easy to find extreme ideological positions on many current issues in ethics and technology – such as genetically modified food or vaccination – but very hard to find out what most people would choose and why.

What is the role for artificial ethics here? Here is a proposal: think of a recent "Turing test" whereby for many of us, news.google beats out the parochial editors of more local "papers." Similarly, many of us evidently rely on google for search and Amazon, Pandora, and Slashdot for book, music, and article recommendations. All are successful implementations of value driven social networking software. Regularly and automatically mining

the Net into bits of value, this is itself the best example of successful artificial ethics, at least in the value subfield.[2]

NERD-II notices how acceptable this automated value help appears to be. We propose automated agents to replace our hand-crafted expert and lay advisors. this has some advantages – moving us investigators farther from a biasing role and gives philosophical ethics a chance to be useful – in a testable way. As our Democracy, Ethics and Genomics project tested NERD-I against focus groups and deliberative democracy, NERD-II will allow us to test different methods of ethics within our open and experimental framework. This is the core assignment of my theoretical graduate seminar next year.

We aim at robust reflective equilibrium. 'Equilibrium' signals our game theoretic roots and aspiration to a result where everyone shares common knowledge of the situation and other agents. 'Reflective' adds a 'normative' dimension: each should be pressed to reflect, answer challenges etc. Finally, we aspire to a robust, dynamic public site, challenged by the unanticipated free actions of a free and creative population armed with the information and social connections available on the Internet.

Conclusion

Each of these three stages had a different take-away for artificial ethics. Formal game theoretic and computational modeling can usefully extend the science of ethics. It can show us what cannot be done, for example, reminding us of the formal limits on rationality, altruism, reciprocity, and social norms. Simulation allows us to play with more complex situations and agents while maintaining a link to theoretical rigor.

To end on a practical note, consider my encounter with robotics. (Danielson 1992) foolishly included 'robots' only in the title, which got the book misclassified by the Library of Congress. (So much for overstating one's results.) The EAME project suggested that one limitation of simulations could be overcome by building communities of robots, so solve problems like "fire in the theatre", which we did in a very fruitful graduate seminar (Danielson 1999). Now working with students in UBC's Animal Welfare Program, we plan to deploy various robotic projects, such as robots to play with cats and robots to test families before they adopt animals from shelters (Danielson 2001b). These robots need to be safe, simple for cats and people to use, and morally acceptable as well. To test the former and drive the latter, we will feature them in the NERDII context, i.e. open to full but ethically structured public evaluation.

In closing, we return to the question of EthicAlife (EA). What does Artifical Ethics have to say about creating machines which engage in our moral life and make ethical judgments?

Nothing except the appeal to experimental evidence above is restricted to human agents. Therefore, the lessons of all three parts of the AM research program apply to attempts to design or evolve EA.

1. EA will need the tool of reciprocity; they will need to distinguish owners, friends, innocents, and foes. The designs will have unintended interactions and will need to be tested against unanticipated interactants, like kids, cats, kibitzers, and evil-doers.
2. With the best intentions, no one gets to design the world, so there will be many kinds of EA, like there are several cooperative types found in human agents. Attempts to certify "genuine" EA will be about as successful as attempts to certify professionals or perhaps limit the clients that are connected on the open Internet.
3. People (and perhaps EA) will need sophisticated tools to evaluate the increasing complex ethical challenges of a world made up of people, complex tools, and perhaps autonomous EA. We hope our own tool, NERD, will play a leading role in meeting such challenges.

NERD-II is process oriented. Roughly, on the question of new entities, it asks the existing ethical agents to evaluate the request to allow new entries to the community[3]. EA needs to be to be evaluated like any other new technology. Indeed NERD was designed to evaluate new genomic technologies, arguably the closest we have come to truly new members of our extended communities. Some may wish to argue that they are creating new persons, which have no more need to justify their addition to the community than do new humans. I will leave this moral or ethical dispute to be played out in an appropriate forum, perhaps NERD-IX.

Acknowledgments

Thanks to the NERD team for making this research possible and fun and to Bill Harms and Rik Blok for work on the EAME project. We are happy to acknowledge long term support from SSHRC for a series of projects (Artificial Morality, Computer Ethics through Thick & Thin (filters), Evolving Artificial Moral Ecologies and Modeling Ethical Mechanisms and generous support from Genome Canada/Genome BC for the Democracy Ethics and Genomics and Building a GE3LS Architecture

[2] Of course, eBay and craigslist belong here too, but their inclusion offends some academics, adverse as our guild can be to markets.

[3] See (Nozick 1974) Chapter 10 for a seminal discussion this process.

projects. NSERC and NSF for support for graduate and post-doctoral students.

References

Ahmad, R., Bornik, Z., Danielson, P., Dowlatabadi, H., Levy, E., Longstaf, H., et al. (2005). *Innovations in web-based public consultation: Is public opinion on genomics influenced by social feedback?* National Centre for e-Social Science, Univ. of Manchester.

Ahmad, R., Bornik, Z., Danielson, P., Dowlatabadi, H., Levy, E., Longstaf, H., et al. (In press). A Web-based Instrument to Model Social Norms: NERD Design and Results. *Integrated Assessment*.

Boyd, R. & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiology*, *13*, 171 - 195.

Danielson, P. (1991). Closing the Compliance Dilemma. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice: Essays on Gauthier*. New York: Cambridge University Press.

Danielson, P. (1992). *Artificial morality: virtuous robots for virtual games*. London: Routledge.

Danielson, P. (1995b). From Rational to Robust: Evolutionary Tests for Cooperative Agents.

Danielson, P. (1995a). Making a Moral Corporation: Artificial Morality Applied. *Financial Accounting 5*. Vancouver: Certified General Accountants" Association.

Danielson, P. (1998b). Evolution and the Social Contract. *The Canadian Journal of Philosophy*, *28*(4), 627 - 652.

Danielson, P. (1998a). Evolutionary Models of Cooperative Mechanisms: Artificial Morality and Genetic Programming. In P. Danielson (Ed.), *Modeling Rationality, Morality, and Evolution 7*. (pp. 423-41). New York: Oxford University Press.

Danielson, P. (1999). Robots for the Rest of Us or for the 'Best' of Us. *Ethics and Information Technology*, *1*, 77 - 83.

Danielson, P. (2001b). Ethics for Robots: Open-ended Ethics & Technology. *1*, 77 - 83.

Danielson, P. (2001a). Which Games Should Constrained Maximizers Play?. In C. Morris, & A. Ripstein (Eds.), *Practical Rationality and Preference: Essays for David Gauthier*. New York: Cambridge University Press.

Danielson, P. (2002). Competition among Cooperators: Altruism and Reciprocity. *Proceedings of the National Academy of Sciences*, *99*, 7237 - 7242.

Danielson, P., Ahmad, R., Bornik, Z., Dowlatabadi, H., & Levy, E. (in press). Deep, Cheap, and Improvable: Dynamic Democratic Norms & the Ethics of Biotechnology. *Journal of Philosophical Research, Special Issue on Ethics and the LIfe Sciences.*

Gauthier, D. (1977). Social Contract as Ideology. *Philosophy and Public Affairs*.

Gauthier, D. (1984). Deterrence, maximization, and rationality. In D. MacLean (Ed.), *The Security Gamble: Deterrence Dilemmas in the Nuclear Age*. (pp. 101 - 22.). Totowa, N.J.: Rowman & Allanheld.

Gauthier, D. (1988b). Moral Artifice. *Canadian Journal of Philosophy*, *18*, 385-418.

Gauthier, D. (1988a). Morality, rational choice, and semantic representation. *Social Theory and Practice*, *5*, 173-221.

Gauthier, D. (1991). Why Contractarianism?. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice*. New York: Cambridge Univ. Press.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge Mass.: MIT Press.

Kurzban, R. & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *PNAS*, *102*(5), 1083-807.

Nozick, R. (1974). *Anarchy, state, and utopia*. New York: Basic Books.

Ostrom, E., & Walker, J. (2003). *Trust and reciprocity : interdisciplinary lessons from experimental research*. New York: Russell Sage Foundation.

Sen, A. (2002). *Rationality and Freedom*. Cambridge, Mass.: Harvard University Press.

Skyrms, B. (1996). *Evolution of the social contract*. Cambridge ; New York: Cambridge University Press.