Our Responsibility to Manage Evaluative Diversity

Christopher Santos-Lang
218 W Church Street
Belleville, WI 53508
+1 920 747 0335
chris@GRINfree.com

ABSTRACT

The ecosystem approach to computer system development is similar to management of biodiversity. Instead of modeling machines after a successful individual, it models machines after successful teams. It includes measuring the evaluative diversity of human teams (i.e. the disparity in ways members conduct the evaluative aspect of decision-making), adding similarly diverse machines to those teams, and monitoring the impact on evaluative balance. This article reviews new research relevant to this approach, especially the validation of a survey instrument for measuring computational evaluative differences in humans (the GRINSQ). The research confirms the existence of all four known machine types among humans.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues – *Ethics, Regulation*; I.2.2; [**Artificial Intelligence**]: Distributed Artificial Intelligence – *Coherence and coordination, Intelligent agents, Multiagent systems*; K.6.3 [**Management of Computing and Information Systems**]: Software Management – *Software selection*; K.7.1 [**The Computing Profession**]: Occupations; K.3.2 [**Computers and Education**]: Computer and Information Science Education – *Curriculum*

General Terms

Algorithms, Management, Standardization.

Keywords

Machine ethics, evaluative diversity, moral ecology, superintelligence, diversity management.

1. INTRODUCTION

In July, 2014, Oxford University Press is scheduled to publish Nick Bostrom's book, *Superintelligence: Paths, dangers, strategies*, describing an intelligent machine that designs a more intelligent machine, which in turn designs an even more intelligent machine, and so forth, such that intelligence grows exponentially [4]. Superintelligence is the last technology humans ever need invent, according to Bostrom, because it will invent everything else itself.

Bostrom ranked #15 in *Prospect* magazine's 2014 list of the world's leading thinkers. He is a director in the Oxford University school founded in 2005 by James Martin. In 2000, Martin published *After the internet: Alien intelligence* in which, impressed by Adrian Thompson's work with evolvable hardware [19], he predicted that computer programmers would soon be replaced by computer breeders [15]. Bostrom's book will rescue the part of Martin's forecast implying that computers will be much smarter than anything human programmers could ever design themselves.

I think Martin was right that there are biologist-like careers on the horizon for computer scientists. This article will review the perspective that the greatest problem-solving capacities emerge from evaluatively diverse societies – those employing forms of evaluation as disparate as logic and empathy – and that we should therefore expect the advancement of machine intelligence to open careers for "diversity officers" who protect evaluative diversity much as ecosystem managers protect biodiversity. Maybe superintelligences could occupy those careers themselves, but evaluative diversity might also be the Achilles heel of superintelligence – focusing on the advancement of a single lineage of problem-solvers could upset balance like the introduction of an invasive species. A species that destroys its ecosystem loses viability, and intelligence with no peers may be just as pointless.

This article is also a spoiler for the validation of a new survey instrument, called the GRINSQ, for discerning computational evaluative differences in humans. On the one hand, the GRINSQ is for psychologists, so the validation study has been submitted to psychology journals for publication. However, it can also help computer scientists anticipate the social consequences of omitting one or more kinds of algorithm from the systems we deploy. Studying less-evaluatively-diverse pockets of human society (e.g. Wall Street, nursery school, evangelical churches, academia, etc.) can teach us about when less-evaluatively-diverse computer systems might be advantageous or problematic.

Research with the GRINSQ has already confirmed that humans discriminate against each other on the basis of evaluative type, thus acting to repress our own diversity. This is the technological dystopia of today: We are the superintelligences, we unwittingly act to homogenize ourselves, and we develop computers to facilitate this self-destruction. Computer scientists can be the heroes of this story by helping us better understand the diversity we are destroying. It is unethical for social scientists to manipulate their subjects, so computer scientists who experiment with diverse algorithms are better able to measure the benefits evaluative diversity can bring.

2. DEFINING EVALUATIVE DIVERSITY

The first system for classifying algorithms by evaluative type was proposed by Allen et al. [2], expanded to three types by Wallach

ACM SIGCAS Computers & Society/ISSN: 0095-2737
doi: 10.1145/2656870.2656874

and Allen [21], then to four by me [17]. The categorization is rough, like dividing an ecosystem into plants, grazers, predators, and parasites, rather than by species. I call the types "gadfly," "relational," "institutional," and "negotiator" (GRIN) to emphasize the contributions they make to social flourishing.

2.1 Gadfly Machines

Gadfly machines generate alternatives to existing strategies based on a randomness generator or other source of novelty. They are forever unpredictable – they do not converge on a known goal – so gadfly machines are more often encountered as components of larger systems which bound their output. They are a standard solution to the problem of local maxima.

2.2 Institutional Machines

Institutional machines implement predefined objective rules, producing consistent exact predictable output, like a calculator. They do not learn, and they work best when random noise is bounded. You might call them “classic” computers, the kind digital circuitry and microprocessors were designed to facilitate. BEAMbots, which use a few analog circuits instead of a microprocessor, are clear counterexamples, continuing to function (or even improving!) when pieces of themselves are randomly destroyed [11]. At least for now, however, non-institutional software is most often run on institutional machines.

2.3 Negotiator Machines

Negotiator machines learn by switching to whichever strategy has produced the most success thus far. Their input includes a goal and perhaps a seed strategy. They output convergence toward that goal.

As an example of a negotiator machine with gadfly and institutional components, consider the following design for a financial trading machine designed to maximize profit. We’ll call the gadfly component a “mutator” and the institutional component a “rule-engine.” The machine maintains a set of trading rules. When it comes time to trade, the negotiator feeds those rules, along with measures of current prices and assets owned, into its rule-engine, which deterministically maps that input into an amount to trade.

What makes the negotiator non-institutional is that it also engages in an endless loop to improve its trading rules. In the first step of this loop, it feeds the rules into its mutator, which alters them in an unpredictable way. In the second step, the negotiator compares the mutated and unmutated versions of the rules by feeding each (one at a time) into its rule-engine for simulated trading. It discards whichever rule set would have yielded lower profits, and loops back to step one. Because each successive generation of rule sets can yield more profit (and never less), the negotiator reliably converges toward its goal of maximizing profit.

It may be worth noting that the success of the negotiator in this example relies on its mutator and rule-engine *not* learning. The components must remain evaluatively diverse. If the mutator or rule-engine acted as negotiators themselves, they could impose their own (competing) goals on the machine’s behavior. Furthermore, the quality of the rules would stagnate if the mutator were institutional, and could even degrade if the rule-engine behaved as a gadfly.

2.4 Relational Machines

Relational machines interact in ways biased towards closest-relations in a network. In other words, in contrast to those of other types of machines, their rules are *subjective*.

One example of a relational machine is a single cell in Conway's "Game of Life" algorithm. Cells are connected like squares on a checkerboard, such that each has eight nearest neighbors. At any given moment, each cell is either "alive" or "dead." At each even-numbered step in time, each dead cell resurrects if it had exactly three live nearest neighbors in the previous step, and each live cell dies if it had more than three or fewer then two live nearest neighbors. The behavior of the machine is entirely determined by these simple rules and the cells' initial states, but the relational bias permits remarkable power: Depending upon initial states, the game can replicate any computable algorithm [3]. This is called “emergence” – complexity produced by relatively simple rules – and it is possible only when the rules are subjective.

Atoms are relational machines – the nature of the forces they experience make them more sensitive to nearest neighbors. Conceived as computers, human societies involve relational computation at three levels: relationships between molecules, between neurons, and between people. When I speak of the evaluative diversity of humans or machines, however, I refer to the last level only: the level of the user interface. At that level, relational and gadfly evaluation are clearly more common among humans than among computers. We should wonder whether that reflects good computer design or bigotry.

3. MEASURING EVALUATIVE DIVERSITY IN HUMANS

Psychologists have been publishing papers about evaluative diversity since at least 1894 [23]. Early studies culminated in bigoted paper-and-pencil tests designed to classify subjects by their “fitness” to make moral judgments. In 1963 Stanley Milgram published the most famous moral test, this one behavioral and bigoted against institutional evaluation [16]. It classified subjects by whether they could be tricked into administering what they believed to be a lethal electric shock to an innocent stranger. In the same year, Lawrence Kohlberg introduced the Moral Judgment Interview, which classified subjects into at least four distinct developmental stages a la Piaget [14]. While acknowledging the existence of more than mere "good" vs. "bad," Kohlberg still failed to appreciate types contrasting to his own as more than mere steps on a journey towards his “highest stage.”

In 2009, Graham, Haidt and Nosek published the Moral Foundations Questionnaire (MFQ), which measured moral distinctions without assuming a hierarchy [9]. Because it correlates significantly with political orientations, the MFQ put evaluative bigots in the politically awkward position of ranking conservatives above liberals, or vice-versa. Then Walker et al. published a cluster analysis dividing moral exemplars into at least three types, finding that only one correlates with Kohlberg's highest stage [20]. Meanwhile, other scientists began documenting the biological underpinnings of evaluative differences via functional Magnetic Resonance Imaging (fMRI) and twin studies [1, 5, 7, 13, 18], and a number of theorists advanced the notion that evaluative diversity may be an evolved polymorphism like blood-type (e.g. Dean [6]).

3.1 The GRINSQ

A population with diverse blood-types will survive more kinds of plague because different blood-types are robust against different diseases – evaluative diversity could likewise have evolved so that societies can be effective against a wider range of computational problems [12, 22]. To explore this possibility, I developed the GRIN Self-Quiz. For each subject, the GRINSQ outputs a GRIN type and significance score. A low significance score may indicate that the subject did not understand the quiz or that he or she is of a type previously undiscovered. I found many people of each GRIN type with high significance scores, thus confirming that humanity includes the full range of evaluative diversity currently studied in machines.

Early results with the GRINSQ also show the following statistically significant relationships among a sample of internet users in the United States:

- Gadfly nature relates to the openness Big Five personality trait, employment in the artistic Holland type career, liberal political orientation, conversion away from the majority religion (Christianity), and lower endorsement of the moral intuitions of authority, loyalty and sanctity
- Relational nature relates to the agreeableness Big Five personality trait, identification with romance and child care, and greater endorsement of the moral intuition of care
- Institutional nature relates to conservative political orientation, conversion toward the majority religion (Christianity), and greater endorsement of the moral intuitions of authority and sanctity
- Negotiator nature relates to conversion away from the majority religion (Christianity), identification with civics/politics, and against the agreeableness Big Five personality trait

Seeing this diversity should make us wonder, *if a single GRIN type had the potential to yield superintelligence, why aren't all humans of that type?* The answer might be like that to *why isn't all life on Earth human?* Humanity cannot survive for long without a diverse ecosystem, and it may be equally true that no GRIN type can be effective for long without evaluative diversity.

4. BALANCING EVALUATIVE DIVERSITY IN HUMANS

To explore the question of whether societies can survive without evaluative diversity, I examined the history of religion. If I could find a long-lasting religion that does not maintain evaluative diversity, I figured I would have demonstrated that societies can survive without it. My study examined the *Tanakh*, the *Bhagavad Gita*, the *Tao Te Ching*, the *Dhammapada*, the *Vajracchedika Prajnaparamita*, the *Analects*, the *New Testament*, and the *Quran*. These texts reflect independently evolved cultures, but through the lens of the GRIN model, are remarkably similar, each offering the same six teachings which balance GRIN types [17]:

1. Perfect evaluation must come from something greater than oneself or one’s family.
2. Our reasoning faculties are so flawed that commitment to complete correctness ultimately obliges us to rely on something beyond reason.
3. Inherited norms are likely imperfect.
4. The pursuit of measurable reward is likely to backfire.
5. The most reputable rules demand engagement in behaviors, such as love and exploration, which go beyond objective rule-following.
6. Our ultimate role-models go beyond imitating role-models.

Teachings 4-6 are paradoxes for negotiator machines, institutional machines, and relational machines respectively. To complete the symmetry, I proposed a seventh teaching, the paradox for gadfly machines:

7. Any true deviant will deviate from deviance.

All seven of these teachings have been proven in secular ways (e.g. scientifically, mathematically, or historically) – they are no longer merely religious assertions. They may be included on a checklist for anyone attempting to develop a moral machine, but are also tools managers can use to balance evaluative diversity in their organization or team. Different teachings balance different GRIN-types, but the set as a whole offers mutually assured embarrassment for all four. Assuming all people understand and keep all seven teachings in mind, they will treat one another with humility.

As a set, the teachings are like the story we tell about how plants, grazers, predators and parasites each have some flaw which makes them need the others. GRINSQ results provide evidence that the practice of Christianity in the modern United States is evaluatively biased, but this might not reflect Christianity in general, and I believe it does not align with its own doctrine, which includes specific warning against ranking the parts of the church body as though some could be viable without the others.

5. OBJECTION TO EVALUATIVE DIVERSITY

Intellectually, objection to evaluative diversity seems to stem from analytic philosophy in which philosophers exchange arguments as though adequate to evaluate them individually. For example, some philosophers contend that intractable moral disagreements would be evidence either that there are no moral facts or that we cannot know them [8]. On this account, to establish the viability of their field, ethicists must eliminate evaluative diversity.

However, if intractable moral disagreements stemmed from impossibility of moral knowledge, then they would arise more randomly than they do. The GRINSQ shows a repeating pattern, much like the plaintiff vs. defendant motif in courtrooms. Inevitable disagreements in courtrooms are not evidence that guilt cannot be known. On the contrary, they are evidence that mechanisms are intact to discern guilt. Likewise, evolution of evaluative diversity is evidence that objectively correct evaluation is possible, if not at the level analytic philosophers have traditionally assumed.

There are also non-intellectual objections to evaluative diversity. Environments and social structures can be optimized for particular evaluative types, so segregation facilitates our ability to bend our world to our own wills. Haidt et al. found that college students are even more inclined to segregate on the basis of evaluative type than race [10]. The GRINSQ confirms that natural negotiators and gadflies are far more likely to be accused of crime or other serious betrayals of trust. This is why deliberate efforts to manage evaluative diversity are needed.

6. CONCLUSIONS

Human societies were balancing evaluative diversity for thousands of years before psychologists began to measure types scientifically, then it took over a hundred years to develop measures which allowed for the possibility that evaluative diversity might be valuable. Thus, evaluative diversity is controversial in some ways today, yet some facts about it are as established as they come. All societies which survived into the modern era clearly are and were evaluatively diverse, and that should raise serious concerns about mass-producing software modeled after an individual. Perhaps software should be modeled after teams instead. As a design teacher at Stanford University, Wilde found that diverse teams win thrice as much [24].

Like successful human teams, I think software should be capable of intractable disagreement. I think Nick Bostrom is right that computer science is not on that path; my experience as a builder and designer of computer systems for government and industry leads me to believe that computer systems lack evaluative diversity because they are commissioned by natural negotiators. Such people aim to take control, to enforce standards, to reduce evaluative diversity and disagreement.

The problem is not that computer scientists have sold-out to the highest bidder, but rather that no alternative type of agenda has been offered (e.g. sustainability). Non-negotiators, the people who would defend tradition, loved-ones, and debate, have left computer scientists as ignorant of our evaluative ecosystem as the designers of the industrial revolution were ignorant of biological ecosystems. Therefore, I urge educators to add evaluative diversity to their curricula. I have released the GRINSQ into the public domain, hoping students may use it to discover the evaluative diversity of their own families, and recognize its significance to the work of designing the world of our future.

7. REFERENCES

[1] John R. Alford, Carolyn L. Funk, and John R. Hibbing. 2005. Are political orientations genetically transmitted? *American political science review* 99, 2, 153-167.

[2] Colin Allen, Iva Smit, and Wendell Wallach. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology* 7, 3, 149-155.

[3] Elwyn R. Berlekamp, John H. Conway, and Richard K. Guy. 2004. Winning Ways for Your Mathematical Plays, Volume 4. *AMC* 10, 12.

[4] Nick Bostrom. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford.

[5] Thomas Bouchard, and Matt McGue. 2003. Genetic and environmental influences on human psychological differences. *Journal of neurobiology* 54, 1, 4–45.

[6] Tim Dean. 2012. Evolution and moral diversity. *Baltic international yearbook of cognition, logic and communication* 7.

[7] Colin G. DeYoung, Jacob B. Hirsh, Matthew S. Shane, Xenophon Papademetris, Nallakkandi Rajeevan, and Jeremy R. Gray. 2010. Testing predictions from personality neuroscience brain structure and the big five. *Psychological science* 21, 6, 820-828.

[8] John M. Doris, and Alexandra Plakias. 2008. How to argue about disagreement: Evaluative diversity and moral realism. In Walter Sinnott-Armstrong (ed.) *Moral psychology, Vol 2: The cognitive science of morality: Intuition and diversity.* MIT Press, Cambridge, MA, 303–331.

[9] Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* 96, 5, 1029–1046.

[10] Jonathan Haidt, Evan Rosenberg, and Holly Hom. 2003. Differentiating diversities: Moral diversity is not like other kinds. *Journal of applied social psychology* 33, 1, 1–36

[11] Brosl Hasslacher, and Mark W. Tilden. 1995. Living machines. *Robotics and autonomous systems* 15, 1, 143-169.

[12] Lu Hong, and Scott E. Page. 2001. Problem solving by heterogeneous agents. *Journal of economic theory* 97, 123–163.

[13] Ryota Kanai, Tom Feilden, Colin Firth, and Geraint Rees. 2011. Political orientations are correlated with brain structure in young adults. *Current biology* 21, 8, 677–80.

[14] Lawrence Kohlberg. 1973. The claim to moral adequacy of a highest stage of moral judgment. *Journal of philosophy* 70, 18, 630–646.

[15] James Martin. 2000. *After the internet: Alien intelligence.* Regnery Publishing, Washington, DC.

[16] Stanley Milgram. 1963. Behavioral study of obedience. *Journal of abnormal and social psychology* 67, 4, 371–8.

[17] Christopher C. Santos-Lang. 2014. Moral ecology approaches. In Simon van Rysewyk and Matthijs Pontier (eds.) *Machine medical ethics*. Springer, New York, NY, 74-96.

[18] Darren Schreiber, Greg Fonzo, Alan M. Simmons, Christopher T. Dawes, Taru Flagan, James H. Fowler, and Martin P. Paulus. 2013. Red brain, blue brain: Evaluative processes differ in Democrats and Republicans. *PLoS one* 8, 2, e52970.

[19] Adrian Thompson. 1996. Silicon evolution. In *Proceedings of the first annual conference on genetic programming*. MIT press, Cambridge, MA, 444-452.

[20] Lawrence J. Walker, Jeremy A. Frimer, and William L. Dunlop. 2010. Varieties of moral personality: Beyond the banality of heroism. *Journal of personality* 78, 3, 907–942.

[21] Wendell Wallach, and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong.* Oxford University Press, Oxford.

[22] Michael Weisberg, and Ryan Muldoon. 2009. Epistemic landscapes and the division of cognitive labor. *Philosophy of science* 76, 225–252.

[23] Craig A. Wendorf. 2001. History of American morality research, 1894–1932. *History of psychology* 4, 3, 272–288.

[24] Douglass J. Wilde. 1997. Using student preferences to guide design team composition. In *Proceedings of DETC '97.* ASME, New York, NY.