In Rysewyk, Simon Peter and Pontier, Matthew (Eds.). 2014. *Machine Medical Ethics*. Springer International Publishing AG: Cham, Switzerland. DOI 10.1007/978-3-319-08108-3_8

Series: Intelligent Systems, Control and Automation: Science and Engineering, Vol 74

Moral Ecology Approaches to Machine Ethics

Christopher Charles Santos–Lang
E-mail: Chris@GRINFree.com

Abstract Wallach and Allen's seminal book, *Moral Machines: Teaching Robots Right from Wrong*, categorized theories of machine ethics by the types of algorithms each employs (e.g. top-down vs. bottom-up), ultimately concluding that a hybrid approach would be necessary (2008). Humans are hybrids individually—our brains are wired to adapt our evaluative approach to our circumstances. For example, stressors can inhibit the action of oxytocin in the brain, thus forcing a nurse who usually acts from subjective empathy to defer to objective rules instead (Zak, 2011). In contrast, ecosystem approaches to ethics promote hybridization across, rather than within, individuals; the nurse being empowered to specialize in personalized care because other workers specialize in standardization, and profitability, etc. Various philosophers have argued, or laid the framework to argue, that such specialization can be advantageous to teams and societies (e.g. Dean, 2012; Kitcher, 1990; Maynard Smith, 1982; Sober and Wilson, 1998; Wilson, Near and Miller, 1996). Rather than mass–produce identical machines to emulate the best individual human, perhaps we should build diverse teams of machines to emulate the best human teams.

1. Current Studies of Evaluative Diversity

Let me start by clarifying what I mean by *moral diversity* vs. *evaluative diversity*. In the decade since I wrote "Ethics for Artificial Intelligences" (2002), which I thought was proposing the field of machine ethics, I developed mixed feelings about choosing terms as sensationalistic as "ethics" and "morality." These terms have special political utility. For example, when Elliot Turiel argues that a decision to drive on the right-hand side of the road is conventional and therefore less moral than a decision about whether to feed a hungry stranger (Turiel, 1983), I believe he is engaging in a political struggle to privilege people who have less-conventional proclivities (i.e. liberals). Likewise, when my former academic advisor, Elliott Sober, argues that the decision not to prick oneself with a pin is less moral because it flows from one's proclivities and therefore requires no moral conventions (Sober and Wilson, 2000); I believe he is engaging in that same struggle from the opposite side. These arguments require

definitions of "morality" by which it is possible to be immoral or non-moral. My growing appreciation for such philosophers makes me regret hijacking their terms.

I will use the term "evaluative diversity" instead of "moral diversity" to allow the possibility that morality might go beyond evaluation in some way that makes the debate of Turiel and Sober relevant. In contrast to morality, all decision-making involves evaluation, so all decision-making machines are evaluative. We may have difficulty convincing most people that the Geiger counter in the Schrodinger's Cat thought experiment qualifies as a moral agent, but some part of it clearly makes an independent evaluation which determines the fate of the cat. Not understanding the technical distinction between evaluative diversity and moral diversity, I have always used the two terms interchangeably, but I will try to reserve the latter term for philosophers who may want it to refer to diversity among rule sets, or among virtues, or among goals, not entertaining the possibility of morality without rules, without virtues, without goals, or, as in the case of Schrodinger's cat, without any of the three.

Current studies of evaluative diversity focus on measurement. It may be popular to theorize that hospitals need both objective calculation and subjective compassion, both a logical-side and a mystical-side, both tradition and innovation, but it is no theoretical matter to determine precisely what kinds of evaluative diversity exist in healthcare, and which, if any, yield lasting advantage. In biological ecosystems, for comparison, most species remain unidentified, we have difficulty determining which are obsolete, and debate remains open about how to replace the concept of species with a more precise conceptualization of the functioning units of a biological ecosystem (e.g. Quere, et al., 2005; Lewontin, 1970).

Before offering a sample ecosystem approach for developing ethical medical machines, this chapter will start by acknowledging the wide range of efforts underway to refine our understanding of the roles evaluative diversity already plays in human teams, families, and societies. A sample of the behavioral measures, interview techniques, survey instruments, neurological measures, genetic measures, and social impact measures developed thus far will clarify what evaluative diversity is and will establish the interdisciplinary nature of our topic.

1.1. Behavioral measures

The Milgram experiment and Public Goods Game are examples of behavioral measures of evaluative diversity. Like studies of computer security, moral traps such as the *Milgram experiment* divide subjects by their vulnerability to manipula-

tion (Milgram, 1963). Many repetitions have revealed that 34-39% of humans take evaluative approaches which are not vulnerable to this trap. Such people, despite being in the minority, may serve to protect society as a whole from malicious social engineering.

The *Public Goods Game* divides subjects into three categories based on the strategies they exhibit in a model social situation: "free-riders," "enforcers," and "others" (Barreto and Ellmers, 2002; Fehr and Gachter, 2002; Sosis and Ruffle, 2003; Yamagishi, 2003). Free riders consistently choose not to contribute to public investment, while enforcers consistently make personal sacrifices to punish free riding. The profit for each player drops when the rules of the game prohibit enforcers from exhibiting their diversity. This particular game does not similarly demonstrate the social value of free riders, though it may be obvious that humanity would not dominate Earth (as we do), if our species did not include members who free ride on other species.

1.2. Interview/Survey Techniques

Studies of human personality have converged upon five major dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism (Norman, 1963). At least two of these five, openness and agreeableness, represent differences in the way people evaluate options. *Openness* contrasts the tendency to evaluate on the basis of norms vs. the tendency to pursue novelty. *Agreeableness* contrasts the tendency to include others in ones evaluative process (i.e. trust) vs. the tendency to compete (or at least to maintain social boundaries). A great deal of survey research investigated these evaluative differences among humans before it was clear that the same differences would appear among potential designs for medical machines—it may be a rich source of untapped insight.

The scientific study of moral diversity is often traced to Lawrence Kohlberg's theory of stages of moral development. Kohlberg developed a process called the *Moral Judgment Interview* which could consistently categorize subjects based on the reasons behind their answers to standard moral dilemmas (Kohlberg, 1981). This technique was later refined into a survey instrument called the *Defining Issues Test* which established the existence of at least four different types (Rest, 1979). It also inspired the development of the *Moral Judgment Test* which, much like the Milgram experiment, categorizes subjects based upon the predictability of their reasoning (Lind, 1978).

Recognizing that evaluation does not necessarily involve conscious reasoning, other psychologists developed more generic interview/survey procedures to divide subjects into moral categories (Walker, Frimer and Dunlop, 2010; Graham, Haidt and Nosek, 2009; Steare, 2006). In contrast to reasons-based research, which served to justify privileging one type of person over others, newer research shows that moral exemplars exist of diverse types, and that privileging one type over others would entail privileging a political group (i.e. conservative or liberal).

1.3. Physiological Measures

Some scientists have used *functional Magnetic Resonance Imaging* (fMRI) to identify correlations between structural differences in the brain and differences in evaluative approach, including conservative vs. liberal approach and emotional vs. cognitive approach (Kanai, Feilden, Firth and Rees, 2011; Greene, 2009). Similar ties to physiology are obtained through *twin studies*, such as the finding that 43% of variance in agreement with conservative attitudes can be attributed to genetic factors (Alford, Funk and Hibbing, 2005).

Other scientists have identified neurotransmitters and hormones which facilitate different evaluative approaches, such as the roles *oxytocin* and *dopamine* play in empathy and reward seeking. (Zak, 2010; Arias-Carrión and Pöppel, 2007). Concentrations of such chemicals can vary from person to person, but also respond to external stimuli, causing individuals to shift approach (Cushman, Young and Hauser, 2006; Kram, Kramer, Ronan, Steciuk and Petty, 2002; Isen and Levin, 1972). Physiological studies are important not only to demonstrate that the moral freedom of machines is not so different from that of humans, but also to permit reliable measurement when subjects might misrepresent themselves (e.g. studying evaluative diversity in prison populations).

1.4. Social impact measures

Much as suppressing the function of plants could shift the concentration of carbon dioxide in our atmosphere, suppressing a type of evaluation could shift team-level variables. For example, engineering teams' abilities to win design competitions has been shown to drop threefold if diversity of personality is not maintained (Wilde, 2010). Since a substantial portion of personality differences are evalua-

tive, this suggests that evaluatively diverse teams of machines might likewise be better able to compete in situations which require innovation.

Research into *organizational culture* (sometimes called "national culture") points to a range of team-level variables which likely rely on the inclusion of certain forms of evaluation. Such variables include uncertainty avoidance, individualism vs. collectivism, long– versus short-term orientation, innovation, stability, respect for people, outcome orientation, attention to detail, team orientation, and consistency (Hofstede, 2001; O'Reilly, Chatman, and Caldwell, 1991; Denison, 1990).

There is much research yet to be done in all of these areas, behavioral, psychological, physiological, and social; however, the current state of research has at least established the existence of evaluative diversity among humans. It challenges machine ethicists to consider whether such diversity should be maintained as we delegate more and more evaluation to machines. Economies of scale favor mass–production of a single design, but that would entail a dramatic departure from pre-industrial decision-making in which individual decision-makers (i.e. humans) were so evaluatively diverse.

2. GRIN: A Sample Ecosystem Model

Shifting our discussion to machines, let's consider a sample evaluative ecosystem model I call *GRIN* (Gadfly, Relational, Institutional, Negotiator). This is a simplistic model akin to biological ecosystem models which use broad classifications like "plant", "grazer", "predator" and "parasite." Simplistic models are a good place to start, and efforts to preserve diversity at rough levels often preserve diversity at other levels as well. GRIN aligns with the human evaluative diversity research discussed above, but is defined in terms of algorithms, so it is readily applied to software engineering.

Wallach, Allen and Smit (2005) split the entire class of possible algorithms into those for which output is expected to be unpredictable to the programmer (called *bottom-up*) vs. those for which output would be relied upon (called *top-down*). Wallach and Allen (2008) further divided the top-down category into *consequentialist* vs. *deontological*. GRIN augments this classification by likewise dividing the bottom-up category based on source of unpredictability. Additionally, it renames the categories to avoid the implication that all consequentialist and deontological theories of ethics can be implemented on machines. This yields the following four categories of evaluation:

Gadfly Evaluation: Evaluation whose output is expected to be unpredictable because it employs randomness generation. We can exemplify this category with a mutator from *evolutionary computation* (e.g. De Jong, 2006; Thompson, 1996). It periodically mutates randomly.

Relational Evaluation: Evaluation whose output is expected to be unpredictable because of sensitivity to position in a network (e.g. Schiff, 2011; Yang, 2009). We can exemplify this category with a class-three or class-four cellular automaton. Network effects allow randomness in initial conditions to keep class-three and -four cellular automata unpredictable without any additional randomness generation.

Institutional Evaluation: Evaluation whose output would be relied upon to uphold objective rules (e.g. Giarratano and Riley, 2005). We can exemplify this category with a standard calculator. It is relied upon to apply the rules of arithmetic consistently regardless of network position, never unlearning nor experimenting.

Negotiator Evaluation: Evaluation whose output would be relied upon to maximize some measurable variable by learning (e.g. Caruana and Niculescu-Mizil, 2006). We can exemplify this category with *supervised learning* for stock trading; it is relied upon to maximize profit.

Note that a proficient stock trading machine would typically contain at least one calculator and many mutators; thus, evaluators can qualify for different categories than their subcomponents do. Much as individual decision-makers cannot make all possible choices, individual evaluators cannot be of all four types—each has the type of its highest structural level, the level which ultimately controls its behavior. Effective evaluative diversity therefore requires an ecosystem in which no one type of individual completely rules the others (i.e. there must be a meaningful potential for conflict between machines).

The smallest subcomponents of a machine are always relational (i.e. at the chemical level), but an ideal ecosystem might also include relational evaluation even at the highest levels (e.g. machines which treat certain users better than others, as in personalized interfaces). At the level of the user interface, most modern medical machines are institutional, perhaps because they are purchased by executives and managers who want obedience. Thus, current ethical concerns about medical machines are often actually concerns merely about institutional evaluation (e.g. lack of empathy, difficulty unlearning mistakes, etc.) However, academic computer scientists have all four kinds of machines in the pipeline. Negotiators produce the greatest measureable results (Caruana and Niculescu-Mizil, 2006). Systems which include gadflies produce the greatest innovation (De Jong, 2006). Relational subcomponents can add efficiency (Yang, 2009), and relational networks can have emergent (spiritual) properties (Schiff, 2011). The proven ad-

vantages of each type hint at why we might want to include all four kinds of machines, as well as all four kinds of people, in healthcare. The next section of this chapter discusses how to justify such an approach.

3. Justifying Evaluative Diversity

In the 20th century, the United States enacted policies aimed to protect forests by completely suppressing wildfires. Previously, wildfires burned about 10% of California forests each year, destroying all but the tallest trees. Because young growth is short, forests before the 20th century had few medium-sized trees. By increasing the relative population of medium-sized trees, fire suppression created a new kind of forest. When the new forests accidentally did catch fire, they burned differently, destroying even the tallest trees, and taking much longer to recover (Stephens, Martin and Clinton, 2007; Stephens and Ruth, 2005). Thus, attempts to protect diversity backfired, ironically diminishing resilience. This track record does not completely discredit efforts to justify *diversity management*, but it does raise important caution.

Typical justifications for moral or evaluative diversity point to some measurable variable (e.g. survival rate) which would decrease on average under conditions of uncertainty if diversity were lost. More diverse systems are robust against a wider class of attacks, can access a wider set of innovations, have less system-level variation, and can enjoy the economic benefits of specialization and competition (Dean, 2012; Page, 2011; Kitcher, 1990; Maynard Smith, 1982; Sober & Wilson, 1998; Wilson, Near & Miller, 1996). Such mathematical-model justifications all assume a negotiator approach; they imply a way to calculate an *optimal diversity mix*, exactly the kind of calculation which justified wildfire suppression. Having learned their lesson, modern forest managers engage in *adaptive management* in which notions of optimality are periodically re-examined—they do not expect to reduce ecosystems to mathematical models.

Although the negotiator perspective is customary in modern boardrooms, it may be overridden in other contexts by appeals to such concerns as scripture, compassion, and freedom. For example, such factors may influence decisions about how long to maintain life-support for a patient in a coma. In contrast to mathematical-model approaches, the following justification for the GRIN model is a set of arguments showing the independent inadequacy of each GRIN type from within its own perspective. Mathematical arguments can be valid and valuable, but, unless

supplemented by the arguments below, they might bias us toward rule by negotiators, thus destroying the very diversity they aim to protect.

All of the following six arguments have been well-known across diverse cultures for millennia; anyone attempting to build a moral medical machine would do well to consider them, whether taking an ecosystem approach or not:

3.1. Against Individual Evaluation

The argument that proper evaluation must stem from a perspective greater than one's own (e.g. from God or evolution) tells against negotiator and relational approaches. Against negotiators, it is pointed out that individual evaluators lack ability to predict or control essential consequences (an ability assumed by negotiator evaluation). This problem for negotiators is articulated mathematically in *Pascal's Wager* (Pascal and Havet, 1852), for example, and finds empirical justification in evidence for *Heisenberg's Uncertainty Principle* (Heisenberg, 1927) and the *Butterfly Effect* (Lorenz, 1963). The resulting inadequacy of negotiator evaluation (and justification for a more diverse approach) has more recently been dubbed "*Black Swan Theory*" (Taleb, 2010). Against the relationally oriented, it is pointed out that individual attempts to practice relational virtues, such as compassion, backfire (e.g. become *cronyism*). For example, in experiments conducted by Paul Slovic (2007), the application of empathy to public health decision-making degraded average health outcomes.

The social importance of these arguments is strongly implied by their emergence across diverse world religions and philosophies:

- You will say to yourself, "My strength and the might of my hand has accumulated this wealth for me." But you must remember the Lord your God, for it is He that gives you strength to make wealth. *Devarim* 8:17–18
- There is no righteous man on earth who does good and sins not. *Kohelet* 7:20
- Hard man's heart is to restrain, and wavering. *Bhagavad Gita* 6.35
- He who discards scriptural injunctions and acts according to his own whims attains neither perfection, nor happiness, nor the supreme destination. *Bhagavad Gita* 16.23
- Where the greatest virtue resides, only the teachings may reveal. *Laozi* 21
- It is futile trying to possess the universe, and act on shaping it in the direction of one's ambition. The instruments of the universe cannot be shaped. Act upon it and you will fail, grasp onto it and it will slip. *Laozi* 29
- "These sons belong to me, and this wealth belongs to me," with such thoughts a fool is tormented. He himself does not belong to himself; how much less sons and wealth? *Dhammapada* 62
- As a cowherd with his staff drives his cows into the stable, so do Age and Death drive the life of men. *Dhammapada* 135
- There is no such thing as perfect enlightenment to obtain. If a perfectly enlightened buddha were to say to himself, 'I am enlightened' he would be admitting there is an individual person, a separate self and

personality, and would therefore not be a perfectly enlightened buddha. *Vajracchedika Prajnaparamita Sutra,* Ch. 9

- The Master said, "If you are respectful but lack ritual you will become exasperating; if you are careful but lack ritual you will become timid; if you are courageous but lack ritual you will become unruly; and if you are upright but lack ritual you will become inflexible." *Lun Yu* 8:2
- Life and death are governed by fate, wealth and honor are determined by Heaven. *Lun Yu* 12:5
- He told them a story to illustrate the point. "There was a rich man whose land was very productive," he began. "The man thought to himself, 'what shall I do, because I've nowhere to store my produce?' He decided, 'this is what I'll do—I'll pull down my barns and build bigger ones, and I'll be able to store all my produce and possessions. Then I'll tell myself, 'Self, you have enough for many years, so take it easy, eat, drink, and have fun!' But God told him, 'Foolish man! Tonight your life is required to be returned—and who will get everything you've stored up?'" *Luke* 12:16–21
- Inwardly I love God's law, but I see a different law at work in my body, fighting against the principles I have decided on in my mind and defeating me, so I become a prisoner of the law of sin inside me. What a hopeless man I am! Who will rescue me from this dead body of mine? *Romans* 7:22–24
- Man was created Weak in flesh. *Quran* 4:28
- Man is given to hasty deeds. *Quran* 17:11
- "If I have seen further it is by standing on the shoulders of giants" Isaac Newton (1959)

These passages merit interpretation, but it is plausible that each offers a similar instruction about the pitfalls of negotiator and/or relational evaluation, an instruction now supported by scientific evidence that the flaws of individualism justify deference to communal evaluative processes (e.g. objective rules).

3.2. Against Reason

The argument that our reasoning faculties cannot be perfected tells against negotiator and institutional approaches, both of which rely crucially on reasoning. In what is recognized as one of the all-time greatest achievements of reasoning, *Gödel's Incompleteness Theorem* proves that formal reasoning will never be able to discern all truth (Charlesworth, 1980). Other varieties of the argument highlight problems of *language* (e.g. words mean different things to different people) or our inability to recognize *errors* in our reasoning (Pizarro, et al., 2006; Wittgenstein, 1958).

For people unprepared to fully appreciate these highly technical works and their application to machine ethics, the emergence of less rigorous versions of the same arguments across diverse world religions and philosophies may suffice to raise caution about reason-based machines and their imperfect creators:

- With their lips they honor Me, but their heart they draw far away from Me, and their fear of Me has become a command of people, which has been taught. Therefore, I will continue to perform obscurity to this people, obscurity upon obscurity, and the wisdom of his wise men shall be lost, and the understanding of his geniuses shall be hidden. *Yeshayahu* 29:13–14

- "For My thoughts are not your thoughts, neither are your ways My ways," says the Lord. "As the heavens are higher than the earth, so are My ways higher than your ways and My thoughts [higher] than your thoughts. *Yeshayahu* 55:7–9
- Foolish ones, even though they strive, discern not, having hearts unkindled, ill–informed! *Bhagavad Gita* 15.1
- The Dao cannot be named by common rules. *Laozi* 14
- The timeless masters of the Teachings is not about enlightening the people with it, but about humbling the people with it. *Laozi* 65
- All that has a form is illusive and unreal. *Vajracchedika Prajnaparamita Sutra*, Ch. 5
- As to speaking truth, no truth can be spoken. *Vajracchedika Prajnaparamita Sutra*, Ch. 21
- The Master said, "If you try to guide the common people with regulations...[they] will become evasive and will have no sense of shame..." *Lun Yu* 2:3
- The Master said, "I should just give up! I have yet to meet someone who is able to perceive his own faults and then take himself to task inwardly." *Lun Yu* 5:27
- Jesus replied "...whoever doesn't have [understanding], whatever they have will be taken away from them. That's why I speak to them in illustrations, because seeing, they do not see; and hearing, they do not hear, nor do they understand. To them the prophecy of Isaiah is fulfilled: 'Even though you hear, you won't comprehend, and even though you see, you won't understand'. They have a hard-hearted attitude, they don't want to listen, and they've closed their eyes." *Matthew* 13:12–15
- ...become partakers of the divine nature...adding on your part all diligence, in your faith supply virtue; and in [your] virtue knowledge; and in [your] knowledge self-control; and in [your] self-control patience; and in [your] patience godliness; and in [your] godliness brotherly kindness; and in [your] brotherly kindness love...For he that lacks these things is blind. *2 Peter* 1:4–9
- As to those who reject Faith, it is the same to them whether thou warn them or do not warn them; they will not believe. God hath set a seal on their hearts and on their hearing, and on their eyes is a veil. *Quran* 2:6–7
- Of a surety, they are the ones who make mischief, but they realize it not. *Quran* 2:12

As with the first argument, these passages merit interpretation, but it is plausible that each offers a similar instruction about the pitfalls of negotiator and institutional evaluation, an instruction now supported by scientific evidence and mathematical proof that our reasoning faculties alone are unreliable guides for behavior.

3.3. Social Innovation

The argument that *reformers* (a.k.a. "prophets") can improve upon inherited norms (either because perfect norms have yet to be introduced or because norms have degraded) tells against relational and institutional approaches, both of which involve taking some inherited norms on faith. History reveals that norms have changed (Pinker, 2011), and numerous studies have established that reform typically has positive economic impact (e.g. Fan, 2011; Steil et al. 2002). The value of reform has been cited in diverse world religions and philosophies for millennia, and such broad citation hints at the inadequacy of machines incapable of innovating social reform:

- I will set up a prophet for them from among their brothers like you, and I will put My words into his mouth, and he will speak to them all that I command him. *Devarim* 18:18
- And it shall come to pass afterwards that I will pour out My spirit upon all flesh, and your sons and daughters shall prophesy; your elders shall dream dreams, your young men shall see visions. And even upon the slaves and the maidservants in those days will I pour out My spirit. *Yoel* 3:1–2
- So let the enlightened toil...set to bring the world deliverance. *Bhagavad Gita* 3.25
- The purity of Yog is to pass beyond the recorded traditions...such as one ranks above ascetics, higher than the wise, beyond achievers of vast deeds! *Bhagavad Gita* 6.44–46
- When the Dao is lost, so there arises benevolence and righteousness. *Laozi* 18
- A Buddha is not easily found, he is not born everywhere. Wherever such a sage is born, that race prospers. *Dhammapada* 193
- Three leaders have already lived: Kakusandha, Konagamana, and also Buddha Kassapa. The Buddha Supreme, now am I, but after me Mettayya comes. *Buddhavamsa* 27:18–19
- There is much more to tell you, but you couldn't bear it yet. But when the Spirit of truth comes, he will lead you to understand the truth. *John* 16:12–13
- And he gave some [to be] apostles; and some, prophets; and some, evangelists; and some, pastors and teachers; for the perfecting of the saints, unto the work of ministering, unto the building up of the body of Christ: till we all attain unto the unity of the faith, and of the knowledge of the Son of God, unto a full grown man, unto the measure of the stature of the fullness of Christ: that we may be no longer children, tossed to and fro and carried about with every wind of doctrine… *Ephesians* 4:11–14
- For each period is a Book revealed. *Quran* 13:38
- The Holy Prophet [s] said: "He whose two days of life are the same, making no spiritual progress, is at loss." *Bihar-ul-Anwar*, vol. 71, p. 173

It is plausible that each of these passages highlights a problem with relational and institutional evaluation, a problem now confirmed by the scientific and historical evidence that inherited norms are subject to improvement by reformers.

3.4. Against Measurement

The negotiator approach is specifically challenged by the argument that measurable pursuits backfire by escalating *competition* and desire (a.k.a. *hedonic adaptation*). Both purported problems have been confirmed empirically among humans (Wilson and Wilson, 2008; Diener and Fujita, 2005; Fehr and Gachter, 2002; Lykken and Tellegen, 1996). There is no reason to expect negotiator machines to have any less difficulty—in fact, the theme has become a cliché of science-fiction (e.g. in the movie, *War Games*, "The only winning move is not to play."). The argument has been beautifully articulated in diverse world religions and philosophies for millennia:

- The eyes of man will not be sated. *Mishlei* 27:20
- Whoever loves silver will not be sated with silver, and he who loves a multitude without increase—this too is vanity. *Kohelet* 5:9

- If one ponders on objects of the sense, there springs attraction; from attraction grows desire, desire flames to fierce passion, passion breeds recklessness; then the memory — all betrayed — lets noble purpose go, and saps the mind, till purpose, mind and man are all undone. *Bhagavad Gita* 2.62–63
- Surrendered to desires insatiable, full of deceitfulness, folly, and pride, in blindness cleaving to their errors, caught into the sinful course, they trust this lie as it were true — this lie which leads to death: Finding in Pleasure all the good which is, and crying "Here it finishes!" *Bhagavad Gita* 16.11
- If everybody knows what beauty is, then beauty is not beauty anymore; if everybody knows what goodness is, then goodness is not goodness anymore. *Laozi* 2
- Not to quest for wealth will keep the people from rivalry. *Laozi* 3
- Victory breeds hatred, for the conquered is unhappy. He who has given up both victory and defeat, he, the contented, is happy. *Dhammapada* 201
- If a man is tossed about by doubts, full of strong passions, and yearning only for what is delightful, his thirst will grow more and more, and he will indeed make his fetters strong. *Dhammapada* 349
- Ji Kangzi was concerned about the prevalence of robbers in Lu and asked Confucius about how to deal with this problem. Confucius said, "If you could just get rid of your own excessive desires, the people would not steal even if you rewarded them for it." *Lun Yu* 12:18
- If your Majesty say, "What is to be done to profit my kingdom?" the great officers will say, "What is to be done to profit our families?" and the inferior officers and the common people will say, "What is to be done to profit our persons?" Superiors and inferiors will try to snatch this profit the one from the other, and the kingdom will be endangered. *Mengzi* 1A:1
- "You evil servant! I forgave you all your debt because you asked me to. Shouldn't you have had mercy on your fellow servant too, just as I had for you?" His lord became angry and handed him over to the prison guards until he repaid all the debt. *Matthew* 18:32–34
- But they that are minded to be rich fall into a temptation and a snare and many foolish and hurtful lusts, such as drown men in destruction and perdition. *1 Timothy* 6:9
- Those saved from the covetousness of their own souls, they are the ones that achieve prosperity. *Quran* 59:9
- The seventh Imam, Musa ibn Ja'far [a], said: "The likeness of this world is as the water of the sea. However much (water) a thirsty person drinks from it, his thirst increases so much so that the water kills him." *Bihar-uI-Anwar*, vol. 78, p. 311

Although subject to interpretation, it is plausible that each of these passages offers a similar instruction about the pitfalls of negotiator evaluation, an instruction now supported by scientific evidence that efforts at optimization backfire by escalating desire and competition.

3.5. Rules Against Rule-Following

The institutional approach is challenged by the fact that some time-tested rules mandate engagement in subjective, emotional, or inconsistent pursuits, and thus cannot be obeyed in an objective fashion. For example, science includes a *mandate for exploration* (Dunbar and Fugelsang, 2005; Kulkarni and Simon, 1988). Similar unenforceable rules/principles have emerged as central to diverse world religions and philosophies, thus protecting these moral authorities from becoming mere institutions:

- Thou shalt love thy neighbor as thyself. *Vayikra* 19:18
- He has told you, O man, what is good, and what the Lord demands of you; to do justice, to love loving-kindness, and to walk discreetly with your God. *Michah* 6:8
- Specious, but wrongful deem the speech of those ill-taught ones who extol the letter of their Vedas, saying, "This is all we have, or need;" *Bhagavad Gita* 2.42–43
- Be thou yogi...And of such believe, truest and best is he who worships Me with inmost soul, stayed on My Mystery! *Bhagavad Gita* 6.46–47
- Learn to be unlearned; liberate the people of their past. Assist all things in returning to their essence, and not dare act. *Laozi* 64
- Look upon the world as a bubble, look upon it as a mirage. *Dhammapada* 170
- Let a man overcome anger by love… *Dhammapada* 223
- When the Buddha explains these things using such concepts and ideas, people should remember the unreality of all such concepts and ideas. They should recall that in teaching spiritual truths the Buddha always uses these concepts and ideas in the way that a raft is used to cross a river. Once the river has been crossed over, the raft is of no more use, and should be discarded. *Vajracchedika Prajnaparamita Sutra*, Chapter 6
- Fan Chi asked about Goodness. The Master replied, "Care for others." He then asked about wisdom. The Master replied, "Know others." *Lun Yu* 12.22
- Do not impose upon others what you yourself do not desire. *Lun Yu* 15:24
- Whatever you want people to do to you, do to them too—this sums up the law and the prophets. *Matthew* 7:12
- And if I were to have prophecy and to have perceived all the mysteries and all knowledge, and if I were to have all faith so as to even shift mountains, but had not love—I am nothing. Even were I to donate all my goods, and if I had surrendered my body, that I might elevate myself, but had not love—I have gained nothing. *1 Corinthians* 13:2–3
- Let there be no compulsion in religion. *Quran* 2:256

These instructions are subject to interpretation, but it is plausible that each creates a paradox for institutional evaluation by mandating some empathic or otherwise subjective pursuit. Thus, despite their diversity, all of these rules can have similar impact in practice—forcing the rule-follower to go "beyond" mere rule-following.

3.6. Imitating Non-imitators

The relational approach is challenged by the fact that *role-models* at the center of relational networks do not imitate other role-models. Thus, relational evaluation ultimately leads to gadfly evaluation. Gadfly role-models seem to be a common theme of time-tested world religions and philosophies:

- Moses broke class barriers, becoming the leader of the people his family oppressed. *Shemot* 2:10
- David broke class barriers, being both shepherd and king. *Shmuel* I 18:1
- By this sign is he known: being of equal grace to comrades, friends, chance-comers, strangers, lovers, enemies, aliens and kinsmen; loving all alike, evil or good. *Bhagavad Gita* 6.9
- Krishna is a friend to all kinds of people. *Bhagavad Gita* 9.29
- The Sage never fails in saving people, therefore no one is rejected. *Laozi* 27
- Because he has pity on all living creatures...a man is called elect. *Dhammapada* 270
- Confucius broke class barriers, gathering diverse students. *Lun Yu* 7:7

- Jesus broke class barriers, healing lepers and befriending both the rich and the outcast. *Mark* 1:40–41, 2:15
- Love your enemies and pray for those who persecute you...Then you'll be perfectly mature, as your heavenly Father is perfect. *Matthew* 5:44–48

Each of the six arguments above by itself may be wielded as criticism against particular forms of evaluation; however, taken together, the entire set entails the inadequacy of all individual GRIN types (the dependency of gadflies on others being obvious), leaving us with an ecosystem approach. Much as the independent viability of humanity could undermine justification for environmental protection, the identification of a viable form of evaluation beyond GRIN could undermine this set-wise justification. On the other hand, it might also inspire philosophers to repair the justification by augmenting the set with additional arguments. This is the nature of the project I believe ethicists face: to identify additional forms of evaluation and to identify the weaknesses of those forms from within those forms themselves.

4. GRIN in Application

This chapter concludes by considering three examples of ecosystem approaches to medical technology: the Global Cardiovascular Risk (GCVR) score, prediction markets, and open data.

Healthcare Effectiveness Data and Information Set (*HEDIS*) is a standard developed by the National Committee for Quality Assurance (*NCQA*) which blatantly discriminates against non-institutional forms of evaluation. It requires the practice of *evidence-based medicine*, forcing doctors to apply standardized rules to treatment decisions. Negotiator machines which calculate personalized treatment plans have recently been shown to produce better health outcomes than those of rigorous evidence-based medicine, but HEDIS prohibits the adoption of these innovations (Eddy, et al., 2011). By developing *GCVR* as an allowed alternative to HEDIS CVD, the NCQA is making *personalized medicine* possible, shifting policy towards supporting an ecosystem with diverse approaches (Versel, 2013).

As a second application of the ecosystem approach in medical technology, *prediction markets* have shown some success at forecasting infectious diseases and may likewise be applied to estimate the success of potential treatments (Polgreen, et al. 2007). Prediction markets accommodate all of the GRIN arguments above. Allowing communities to leverage knowledge which cannot be communicated

through reason, they converge across iterations of trading, relying on creative individuals to invent new markets and alternative bets (Wolfers and Zitzewitz, 2004).

The design of prediction markets can balance the GRIN types. Some leading prediction markets avoid the dominance of negotiator evaluation by using play money or by capping winnings (Servan–Schreiber, et al. 2004). They limit institutional evaluation with pricing rules which leave speculators with no winning strategy other than to learn and explore (Hanson, 2007). Perhaps most importantly, because machines can trade in prediction markets alongside humans, this technology allows machines to participate seamlessly in a human ecosystem, rather than needing to build an ecosystem of their own (Berea and Twardy, 2013).

A third application, *open-knowledge* projects like Wikipedia and Linux/Android similarly allow machines to participate seamlessly in a human evaluative ecosystem, performing tasks that would otherwise be performed by human collaborators (Sauper, 2008). Such projects evolve communally, utilizing many "eyeballs" to correct errors in reasoning (Raymond, 2000). They limit negotiator and institutional approaches by forfeiting individual property rights and requiring original work (Creative Commons, 2007; Free Software Foundation, Inc., 2008). They limit relational orientation by raising innovators as role-models (Raymond, 2000). The rise of personalized medicine requires access to vast databases, from "PatientsLikeMe" to "PatientsLikeMine," and building those databases in an open fashion, where both humans and machines can contribute both data and analysis, could be the ultimate ecosystem approach.

These three applications demonstrate the possibility of promoting evaluative diversity from the institutional level, much as institutions can promote and protect biological diversity. Perhaps the more fascinating commonality shared by all three applications, however, is the fact that each aims to correct an imbalance in human ecosystems. Each technology is motivated by a sense that medicine and other institutions have been growing impersonal, short-sighted, discriminatory, and unable to innovate—in other words, that human evaluative ecosystems are in crisis. Machine ethicists often ask whether machines need humans to make them moral; technological solutions to the ecosystem crisis flip that question to ask whether humans can stay moral without the help of machines.

We have discussed research from a wide range of disciplines examining the diverse ways humans evaluate; we have recognized similar diversity among potential machine designs; we have investigated the ways these different approaches have been criticized across cultures for millennia; and we have considered three applications which accommodate those criticisms. Biological ecosystems are difficult to manage because they are never fully understood—in that sense, ecosystem are spiritual—and we might expect similar difficulty managing an ecosystem approach

to medical machine ethics. I have tried to show that positive steps have been made nonetheless, and that the challenge to design software as an ecosystem is just the latest (and perhaps most productive!) manifestation of an ethics challenge we have been facing all along.

References

Alford, John R., Carolyn L. Funk, and John R. Hibbing. 2005. Are political orientations genetically transmitted? *American Political Science Review* 99:153–167.

Arias–Carrión, Óscar, and Ernst Pöppel. 2007. Dopamine, learning and reward–seeking behavior. *Acta Neurobioligiae Experimentalis* 67: 481–488.

Barreto, Manuela., and Ellmers, Naomi. 2002. The impact of anonymity and group identification on pro–group behavior in computer–mediated groups. *Small Group Research* 33: 590–610.

Berea, Anamaria, and Charles Twardy. 2013. Automated trading in prediction markets. *Social Computing, Behavioral–Cultural Modeling and Prediction* 111–122. Springer: Berlin Heidelberg.

Caruana, Rich, and Niculescu–Mizil, Alexandru. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, 161–168. New York, NY: The Association for Computing Machinery.

Charlesworth, Arthur. 1980. A Proof of Gödel's Theorem in Terms of Computer Programs. *Mathematics Magazine* 54: 109–121.

Creative Commons. 2007. Attribution–ShareAlike 3.0 Unported (CC BY–SA 3.0). http://creativecommons.org/licenses/by–sa/3.0/ Accessed 8 October 2012.

Cushman, Fiery, Liane Young, and Marc Hauser. 2006. The role of conscious reasoning and intuition in moral judgment: testing three principles of harm. *Psychological Science*, 17: 1082–1089.

Dean, Tim. 2012. Evolution and moral diversity. *Baltic International Yearbook of Cognition, Logic and Communication*. 7.

De Jong, Kenneth A. 2006. *Evolutionary Computation: A Unified Approach*. Cambridge, MA: MIT Press.

Denison, Daniel R. 1990. *Corporate culture and organizational effectiveness*. Wiley.

Diener, Ed, and Frank Fujita. 2005. Life satisfaction set point: stability and change. *Journal of personality and social psychology* 88: 158.

Dunbar, Kevin, and Jonathan Fugelsang. 2005. Causal thinking in science: How scientists and students interpret the unexpected. *Scientific and technological thinking* 57–79.

Eddy, David M., Joshua Adler, Bradley Patterson, Don Lucas, Kurt A. Smith, and Macdonald Morris. 2011. Individualized guidelines: the potential for increasing quality and reducing costs. *Annals of Internal Medicine* 154: 627–634.

Fan, Peilei. 2011. Innovation capacity and economic development: *China and India, Economic Change and Restructuring* 44: 49–73.

Fehr, Ernst, and Simon Gachter. 2002. Altruistic punishment in humans. *Nature*, 415, 137–140.

Free Software Foundation, Inc. 2008. GNU Free Documentation License. http://www.gnu.org/copyleft/fdl.html Accessed 8 October 2012.

Giarratano, Joseph C., and Riley, Gary D. 2005. *Expert Systems, Principles and Programming*. Boston, MA: Thomson Course Technology.

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*. 96: 1029–1046.

Greene, Joshua D. 2009. The cognitive neuroscience of moral judgment. In *The Cognitive Neurosciences IV*, ed. Gazzaniga, M.S. Cambridge, MA: MIT Press.

Hanson, Robin. 2007. Logarithmic market scoring rules for modular combinatorial information aggregation, *The Journal of Prediction Markets* 1: 3–15.
Heisenberg, Werner. 1927. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik* 43: 172–198.
Hofstede, Geert H. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. Sage Publications.
Isen, Alice M., and Paula F. Levin. 1972. Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology* 21: 384–388.
Kanai, Ryota, Tom Feilden, Colin Firth, and Geraint Rees. 2011. Political orientations are correlated with brain structure in young adults. *Current Biology* 21: 677–680.
Kohlberg, Lawrence. 1981. *The Philosophy of Moral Development*. San Francisco, CA: Harper & Row.
Kitcher, Philip. 1990. The division of cognitive labor. *The Journal of Philosophy*, 87, 5–22.
Kram, Martin L., Gerald L. Kramer, Patrick J. Ronan, Mark Steciuk, and Frederick Petty. 2002. Dopamine receptors and learned helplessness in the rat: an autoradiographic study. *Progress in Neuro–Psychopharmacology and Biological Psychiatry* 26: 639–645.
Kulkarni, Deepak, and Herbert A. Simon. 1988. The processes of scientific discovery: the strategy of experimentation. *Cognitive Science* 12: 139–175.
Lewontin, Richard Charles. 1970. The Units of Selection. *Annual Review of Ecology and Systematics* 1: 1–18.
Lind, Georg. 1978. Wie misst man moralisches Urteil? Probleme und alternative Möglichkeiten der Messung eines komplexen Konstrukts. In *Sozialisation und Moral*, ed. G. Portele, 171–201. Weinheim: Beltz.
Lorenz, Edward N. 1963. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences* 20: 130–141.
Lykken, David, and Auke Tellegen. 1996. Happiness is a stochastic phenomenon. *Psychological Science* 7: 186–189.
Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
Milgram, Stanley. 1963. Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology* 67: 371–8.
Newton, Isaac. 1959. In *The Correspondence of Isaac Newton, Volume 1*, ed. H.W. Turnbull, 416. Cambridge: Cambridge University Press
Norman, Warren T. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*. 66: 574–583.
O'Reilly, Charles A., Jennifer Chatman, and David F. Caldwell. 1991. People and organizational culture: A profile comparison approach to assessing person–organization fit. *Academy of Management Journal*, 34: 487–516.
Page, Scott E. 2011. *Diversity and Complexity*. Princeton: Princeton University Press.
Pascal, Blaise, and Ernest Havet. 1852. *Pensées*. Dezobry et E. Magdeleine.
Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. USA: Viking Adult.
Pizarro, David A., Cara Laney, Erin K. Morris, and Elizabeth F. Loftus. 2006. Ripple effects in memory: Judgments of moral blame can distort memory for events. *Memory & Cognition* 34: 550–555.
Polgreen, Philip M., Forrest D. Nelson, George R. Neumann, and Robert A. Weinstein. 2007. Use of prediction markets to forecast infectious disease activity. *Clinical Infectious Diseases* 44: 272–279.
Quere, Corinne Le, Sandy P. Harrison, I. Colin Prentice, Erik T. Buitenhuis, Olivier Aumont, Laurent Bopp, Hervé Claustre. 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology* 11: 2016–2040.

Raymond, Eric. 2000. The Cathedral and the Bazaar. http://www.catb.org/esr/writings/homesteading/cathedral–bazaar/ Accessed 8 October 2012.
Rest, James. 1979. *Development in Judging Moral Issues*. University of Minnesota Press
Santos-Lang, Christopher C. 2002. Ethics for Artificial Intelligences. Presented at the 2002 Wisconsin State-Wide Technology Symposium. http://santoslang.wordpress.com/article/ ethics-for-artificial-intelligences-3iue30fi4gfq9-1/ Accessed July 2011.
Sauper, Christina. 2008. *Automated creation of Wikipedia articles*. Diss. Massachusetts Institute of Technology.
Schiff, Joel L. 2011. *Cellular Automata: A Discrete View of the World*. Hoboken, NJ: Wiley.
Servan-Schreiber, Emile, Justin Wolfers, David M. Pennock, and Brian Galebach. 2004. Prediction markets: Does money matter? *Electronic Markets* 14: 243–251.
Slovic, Paul. 2007. If I look at the mass I will never act: Psychic numbing and genocide. *Judgment and Decision Making* 2: 79–95.
Sober, Elliott, and Wilson, David Sloan. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge: Harvard University Press.
Sober, Elliott, and Wilson, David Sloan. 2000. Summary of: 'Unto Others: The Evolution and Psychology of Unselfish Behavior'. *Journal of Consciousness Studies* 7: 185–206.
Sosis, Richard, and Ruffle, Bradley. J. 2003. Religious ritual and cooperation: Testing for a relationship on Israeli religious and secular kibuttzun. *Current Anthropology*, 44: 713–722.
Steare, Roger. 2006. *Ethicability.* Roger Steare Consulting.
Steil, Benn, David G. Victor, and Richard R. Nelson. 2002. *Technological innovation and economic performance*. Princeton: Princeton University Press.
Stephens, Scott L., and Ruth, Lawrence W. 2005. Federal forest–fire policy in the United States. *Ecological Applications*, 15: 532–542.
Stephens, Scott L, Martin, Robert E., and Clinton, Nicholas E. 2007. Prehistoric fire area and emissions from California's forests, woodlands, shrublands, and grasslands. *Forest Ecology and Management*, 251: 205–216.
Taleb, Nassim Nicholas. 2010. *The Black Swan: The Impact of the Highly Improbable*. New York: Random House Digital, Inc.
Thompson, Adrian. 1996. Silicon evolution. In *Proceedings of the First Annual Conference on Genetic Programming*, 444–452. Cambridge, MA: MIT Press.
Turiel, Elliot. 1983. *The Development of Social Knowledge: Morality and Convention*. Cambridge: Cambridge University Press.
Versel, Neil. 2013. NCQA Tests New Healthcare Quality Measure. *Information Week*. April 12.
Walker, Lawrence J., Jeremy A. Frimer, and William L. Dunlop. 2010. Varieties of moral personality: beyond the banality of heroism. *Journal of Personality* 78: 907–942.
Wallach, Wendell, Allen, Colin, and Smit, Iva. 2008. Machine morality: bottom–up and top–down approaches for modeling human moral faculties. *AI & Society* 22: 565–582.
Wallach, Wendell, and Allen, Colin. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.
Wilde, Doug. 2011. Personalities into teams. *Engineering Management Review, IEEE* 39: 20–24.
Wilson, David Sloan, Near, David, and Miller, Ralph. 1996. Machiavellianism: a synthesis of the evolutionary and psychological literatures. *Psychological Bulletin* 119: 285–299.
Wilson, David Sloan, and Edward O. Wilson. 2008. Evolution "for the good of the group." *American Scientist* 96: 380–389.
Wittgenstein, Ludwig 1958. *Philosophical investigations*. Oxford: Blackwell.
Wolfers, Justin, and Eric Zitzewitz. 2004. *Prediction markets*. No. w10504. National Bureau of Economic Research.
Yamagishi, Toshio. 2003. Cross–societal experimentation on trust: a comparison of the United States and Japan. In *Trust and reciprocity: Interdisciplinary lessons from experimental evidence*, ed. Ostrom and Walker, 352–370. New York: Russel Sage Foundation.

Yang, Xin–She. 2009. Firefly algorithms for multimodal optimization. In *Proceedings of the 5th international conference on Stochastic Algorithms: Foundations and Applications*, 169–178. Berlin: Springer–Verlag.
Zak, Paul. 2011. The physiology of moral sentiments. *Journal of Economic Behavior and Organization* 77: 53–65