Journal of Evolution & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

25(2) – November 2015

Reframing Ethical Theory, Pedagogy, and Legislation to Bias Open Source AGI Towards Friendliness and Wisdom

John Gray Cox
College of the Atlantic
gray@coa.edu

Journal of Evolution and Technology - Vol. 25 Issue 2 – November 2015 - pgs 39-54

Abstract

Hopes for biasing the odds towards the development of AGI that is human-friendly depend on finding and employing ethical theories and practices that can be incorporated successfully in the construction, programming and/or developmental growth, education and mature life world of future AGI. Mainstream ethical theories are ill-adapted for this purpose because of their mono-logical decision procedures which aim at "Golden rule" style principles and judgments which are objective in the sense of being universal and absolute. A much more helpful framework for ethics is provided by a dialogical approach using conflict resolution and negotiation methods, a "Rainbow rule" approach to diversity, and a notion of objectivity as emergent impartiality. This conflict resolution approach will also improve our chances in dealing with two other problems related to the "Friendly AI" problem, the difficulty of programming AI to be not merely smarter but genuinely wiser and the dilemmas that arise in considering whether AGIs will be Friendly to humans out of mere partisanship or out of genuine intent to promote the Good. While these issues are challenging, a strategy for pursuing and promoting research on them can be articulated and basic legislation and corporate policies can be adopted to encourage their development as part of the project of biasing the odds in favor of Friendly and Wise AGI.

Introduction

Very significant existential risks for humans and other species are presented by possible scenarios for the development of an advanced Artificial General Intelligence (AGI) system that surpasses human intelligence and then begins, perhaps at an exponential rate, to surpass itself (Armstrong 2014, Barrat 2013, Bostrom 2014, Yudkowsky 2008). These scenarios merit not only theoretical consideration but at least the attempt to take precautionary practical action. But there are many difficulties in predicting when and how such scenarios might occur and even conceptualizing clearly what they might consist in and

mean. In the face of these difficulties, Ben Goertzel and Joel Pitt have proposed that: "Without denying the presence of a certain irreducible uncertainty in such matters, it is still sensible to explore ways of *biasing the odds* in a favorable way such that newly created AI systems are significantly more likely than not to be Friendly" (Goertzel and Pitt 2012, 116). The focus of this paper is on the task of formulating the ethical theory, practice and pedagogy that would be appropriate for framing strategies for biasing the odds towards Friendliness.

A central challenge is that mainstream ethical theory and teaching practice seem very ill adapted to provide the content and pedagogy required to develop AGI that will be Friendly in some desirable sense. In so far as they are expressible in algorithmic forms, mainstream theories often seem to provide inconsistent sets of principles that could be programmed only, if at all, in extremely ad hoc ways that remain intuition dependent.[1] And the underlying implication which any superintelligent student of such ethics would be likely to draw is, arguably, that some form of Relativism is true. Such a student would also likely conclude that there are no rational grounds for being biased towards a Friendly attitude and practice with regard to humans. As MacIntyre has argued, the reasons for this Relativist problematic are not trivial (Macintyre 2007). It results from central assumptions and formative intellectual moments in the development of Western philosophy.

The central idea in the solution proposed in this paper is to change the framing of ethics in three ways: 1.) Instead of supposing ethics should provide a monological decision procedure, the proposal is to view it as a dialogical process of negotiation, group problem solving and conflict resolution. 2.) Instead of grounding ethical analysis in a "golden rule" that advocates some form of universal principle (as interpreted in either Utilitarian or Kantian ways), the proposal is to ground ethics in a principle of diversity, a kind of "rainbow rule" that advocates "doing unto others as they would have us do unto them". 3.) Objectivity as a guiding value for ethics should be interpreted not as a form of absoluteness in universality but as a form of impartiality defined in terms of greater inclusiveness of parts in the whole and integrity, resiliency and beauty in developing right relationships between them. These three shifts in the framing of ethics provide a basis for developing hardware, software, learning environments and socio-political contexts for emergent AGIs that could help bias them towards Friendliness.

Three problems with current mainstream ethics – and an alternative framework for ethics

"Mainstream ethics" could be defined ostensibly by pointing to what has been one of the most popular of all courses taught at Harvard in recent years, offered by a highly respected philosopher and teacher, Michael Sandel (Sandel 2009). His course on "Justice" is typical of university classes in that it introduces students to ethical theories through applications to difficult cases in which significant – often life and death – decisions must be made. The aim is to help students refine their understanding of the principles like Kant's Categorical Imperative (CI) and Bentham's Greatest Happiness Principle (GHP). This understanding should include seeing how the principles determine which choices should be made in specific cases and then reflecting on whether those choices reflect students' intuitions about what it would in fact be good to do. Typically the cases are framed so that the choice posed is a dilemma, one in which one choice would be dictated by CI and the other by GHP, thus clarifying their differences and making it possible to assess their merits as theories. For example, students may be asked to suppose they are on a train platform, a trolley car is approaching at great speed, and there is a group of innocent people who will be killed by collision – but they can be saved if the student flips a switch to redirect the train to another track where, however, there is unfortunately a single other innocent person who will be killed by collision. Should the student flip the switch and sacrifice the one to save the many?

For philosophers such dilemmas can be especially interesting because they can serve to sharpen intuitions and concepts – and for teachers they provide exercises that can strengthen student's skills in critical reasoning. But this dilemma based pedagogy reveals three kinds of problems that plague mainstream

ethics and also make it a very poor candidate for framing the development of Friendly AGI. These concern: 1.) its model of reasoning, 2.) its strategies for interpreting the Golden Rule tradition in ethics, and 3.) its conception of objectivity.

Models of Reasoning

The first set of problems are associated with the paradigm of reasoning adopted which is modeled loosely on mathematics and natural science in that it is supposed that given one or more basic axioms, like the CI or GHP, and some specific conditions of the case, a single person acting as a judge or agent can infer what the correct choice would be. As Abney puts it in characterizing Utilitarian, Kantian and other moral systems based on rules or principles: "all rule-based approaches have assumed: (a) the rule(s) would amount to a decision procedure for determining what the right actions was in any particular case; and (b) the rule(s) would be stated in such terms that any non-virtuous person could understand and apply it (them) correctly" (Abney 2012, 36). Such an approach is "monological" precisely in the sense that it assumes that given the principles and specific conditions, one person can determine what is the ethical thing to do. No dialogue is necessary. From a programming point of view, it might seem to be an attractive model of ethics because it suggests that an AGI might be made to be ethical if we can simply assure that its algorithms calculate choices always in accordance with correct assessments of the facts and reliance on the right fundamental principle or principles.

However, one difficulty is that candidates for such fundamental principles like CI and GHP are notoriously ambiguous and difficult for humans to apply in ways that square with their own moral intuitions and that could be modeled in anything other than very ad hoc ways. Such difficulty of application makes them suspect as principles. This suspicion is exacerbated by the highly controversial and unsettled results of attempts to justify or ground them as moral theories. As Abney notes, in considering the search for a unifying and grounded decision procedure in ethics, "despite centuries of work by moral philosophers, no (plausible) such set of rules has been found" (Abney 2012, 37). For many students, these suspicions, coupled with their frustrations in trying to deal with profoundly vexing dilemmas, lead to one of two results: a fideist insistence on some particular dogma or a skeptical/relativist stance. They either dig in their heels, intellectually, proclaim themselves in favor of one theory or the other and attempt to find ad hoc ways to live on its basis or they conclude that in ethics, unlike, perhaps, in science, there are no secure results and no real truths, just personal choices and opinions.

A clue to an alternative way of conceptualizing ethics and its pedagogy is provided by one of the episodes of Sandel's course which is broadcast on Youtube (Sandel 2009). In the dilemma it presents to students, they are asked to suppose they are a doctor with five patients, four of whom are each in need of a new vital organ to survive. The fifth is a perfectly healthy patient who is currently unconscious. The doctor is asked to consider harvesting organs from the healthy patient to save the others. The pedagogical point here is, in part, to push students who had been favoring a strict Utilitarian ethic to consider if their intuitions would indeed lead them to promote the greatest happiness in this case by sacrificing the one for the many. Interestingly, one student proposes an alternative answer. He says he would sacrifice one of the four patients already compromised to harvest organs for the other three – since that victim would have died anyway, given the suppositions of the case. The flow of Sandel's pedagogical process is briefly disrupted by this suggestion. He notes that it is an interesting idea but it spoils the point of his philosophical example. In saying this, Sandel is, of course, following a standard approach in mainstream ethical pedagogy of pushing students to confront the hard cases through maintaining their formulations as dilemmas. This is also the approach that was used by Lawrence Kohlberg in the observation procedures he used to develop one of the most prominent models of ethical development in young people (Kohlberg 1981).

But think about this for a moment. If one of your students comes up with a way of making an ethical dilemma go away, isn't that precisely the sort of thinking you would want to encourage? When he leaves Harvard and acquires a position of some power in the world, the last thing you want is for him to waste time agonizing over forced choices if there is some relatively straightforward way in which he could alter the situation in ways that make the ethically problematic conditions diminish or disappear. The strategy Sandel's student adopted is one that people researching negotiation and problem solving refer to sometimes refer to as "multiplying the options". It is one of a number that can be used to transform the way choice situations are viewed in order to provide better results. This different kind of thinking is one that has been researched and developed extensively over the last forty years in studies of group problem solving, negotiation, mediation, alternative dispute resolution, and conflict management, resolution and transformation.

As a field of study, this one has achieved all the characteristics of maturity. It has standard survey texts such as *Contemporary Conflict Resolution* (Ramsbotham et. al. 2011) and *Peacemaking: From Practice to Theory* (Nan 2011). It has professional journals such as *The Negotiation Journal*, *The Journal of Conflict Resolution*, and *The Journal of Peace Research.* It has centers for research such as the Harvard Negotiation Project and the School for Conflict Analysis and Resolution at George Mason University. It has professional organizations of practitioners such as the Association for Conflict Resolution and the National Association for Community Mediation.

One of the classic texts in the field is *Getting to Yes: Negotiating Agreement Without Giving* (Fisher et. al. 2011). Its authors describe a rich variety of negotiation tactics which grow out of four core strategies: 1.) focusing on the underlying interests behind peoples' explicit positions, 2.) inventing options for mutual gain that "increase the size of the pie", 3.) using objective criteria to assess these, standards that are independent of individual will and can be researched, and 4.) separating out the symbolic and relationship parts of the conflict (the "people part") from the other outcome and action issues and dealing with each appropriately. Principles such as these are used to negotiate conflicts in a wide variety of contexts – from renter/landlord disputes and divorces to labor/management disputes and international treaties such as the Camp David Accords that brokered peace between Israel and Egypt in the Sinai. While successful in many such contexts, the tradition of practice represented by *Getting to Yes* has been criticized as placing too much emphasis on Western and, specifically North American, ways of framing conflicts and dealing with differences. In recent years the study of other important, useful traditions of practice drawing from other cultures has very significantly enriched theoretical and practical insight into ways of dealing with differences. John Paul Lederach's *Preparing for Peac*e (Lederach 1996) provides a very useful introduction to challenges and strategies for developing cross-cultural approaches to dealing with conflicts and the text of Nan et. al. as well as Pat K. Chew's *The Conflict and Culture Reader* (Chew 2001) provide excellent examples of the rich variety of successful traditions that are available.

For our purposes here, two key features of these many varied practices of dealing with disputes should be focused on. First, the practices make central use of a mode of reasoning which is not monological but dialogical. The aim is to talk and interact back and forth with others until some form of shared consent is achieved. The outcome cannot be defined ahead of time by any single party to the dispute. In this way it is quite different form the model of ethical judgment that typifies mainstream ethics in which the courtroom and, in particular, the decision of a judge, is viewed as the paradigm. Instead of a judicial decision procedure for a single judge, a collective negotiation is undertaken. All must share in the process sufficiently to enable them collectively to get to agreement and arrive at the "Yes" of consent.

Second, while the process can make use of insights from formal game theory, it cannot be formulated as a decision procedure coded in an algorithm that would be calculated by one "player" or party to the dispute. This is because these practices of dealing with disputes all emphasize creative initiatives in which the terms of the conflict are redefined and transformed – by revising participants understandings of what their

real interests are, what options may be available, what criteria might be appropriate for assessing them, et cetera. Like the student who disrupted Sandel's neat philosophical dilemma, these traditions of negotiation, mediation and conflict resolution seek to reject dilemmas and transform the conflicts. It is crucial to note here that this does not mean that these practices could not be learned by an AGI. Quite the contrary. The creative activities involved are ones that are teachable and involve using strategies like brainstorming and metaphorical thinking which computers could learn – given the appropriate initial programming and subsequent learning environments. But they must draw on dialogical understandings of reasoning and methods of social research that are grounded in ethnographic understanding and the interpretation of meaning in the context of communities that have developed practices embedded in a life world (Fay 1975, Taylor 1971, MacIntyre 2007). And they must enrich their instrumentalist theories of decision and action with others that include activities framed as expressions, projects and practices in which there are organic relations between means and ends and emergent meanings and values (Cox 1986).

An AGI that is programmed and nurtured in such ways would be arguably more likely to be Friendly to humans for the following reasons. First, it would be more likely to take us seriously as Others, as Persons worthy of dialogue – precisely because it would be more likely to view interaction with us not from the point of view of a monologue with itself about achieving its own goals in a world of scarce and limited resources but from the point of view of a dialogue with us about ways of "increasing the size of the pie" and reaching an agreement in which we as well as it could affirm the "Yes" of consent. Second, such an AGI would be good at – and seek to get continually better at – finding "win/win" kinds of solutions that would enable it to pursue its interests by advancing ours as well. There are no a priori guarantees that the pursuit of solutions will meet with success but promoting the search would bias the odds in favor of the kinds of futures we should be seeking.

Of course this also would not necessarily prevent AGIs from developing interests or ideologies that would lead them to become unfriendly – perhaps a bit like Hitler negotiating with Chamberlain over Poland or, in the classic example of "Realist" politics from Thucydides, like the Athenians negotiating in the "Melian Dialogue". They demanded total capitulation from the weakly defended people of Melos and instead of justifying their demand with any moral claim simply argued "the strong take what they can and the weak suffer what they must" (Thucydides 1910, 5.89.1). This raises the question as to whether there might be some core principle like the Golden Rule which we might seek to encode or nurture in AGIs.

The Golden and Rainbow Rules

This brings us to a second set of problems which mainstream ethics presents for dealing with the Friendly AI problem. The Utilitarian and Kantian principles that characterize the most prominent theories in the field have each been argued to be, in their own way, an interpretation of the Golden Rule inherited from the Christian religious tradition and prominent in various versions in others. That principle can be stated as: "Do unto others as you would have them do unto you." On the face of it it seems reasonable as a principle of ethics, in part because so familiar. And certainly in some specific contexts it seems very appropriate as a principle. In kindergarten one child taking another's toy may be asked to consider "what if Johnny did that to you"; in a market setting one business person thinking of lying to another or breaking a promise might be asked the same sort of thing.

The sense that these acts of stealing or lying would be wrong can be formulated in Utilitarian terms by leading the child or businessperson to see that they would not be happy if stolen from or lied to – and that actions that make everyone happy seem desirable. Or the sense of the wrongness here could also be formulated by leading them to see that they would not feel respected and think it fair if they were stolen from or lied to – and that ethical actions should be ones that can be willed from the point of view of the

recipients of the actions as well as the agents as some version of a Kantian Categorical Imperative would demand.

Part of what makes the Golden Rule and its Utilitarian or Kantian theoretical formulations seem intuitively appropriate in these cases is that the agents and recipients of action in these cases are members of a relatively homogeneous population of kindergartners or businesspeople with similar interests and outlooks.[3] But when we shift to situations in which the recipient of the action is dramatically different, then lots of puzzles emerge. What would it mean, for example, for a grandparent to treat a child as she would have the child treat her? Or how should the Golden Rule be applied when a businesswoman from Germany is trying to deal with a veiled woman in Afghanistan? The Golden Rule would seem to invite ego-centric and ethnocentric behavior that ignores the often very legitimate values and points of view of people from other conditions in life or cultural traditions – just as so many nineteenth century missionaries seemed to have done in promulgating the Golden Rule and the Christian Faith throughout the world.

Of course it might be argued that those missionaries simply mis-applied the rule and that it can work perfectly well in all these cases so long as I make sure, in applying it, to ask not what I, as, for example, a grandparent, might want if I were in the child's situation but what I, as a child, would want in the child's situation. But a key difficulty is that very often, probably typically, it is very difficult for me to imagine what it would be like to really be this other person. The experience of thinking I know what the Other thinks and wants and then finding out I was quite mistaken is not merely common; it is probably one of the definitive experiences of the mature ethical agent who has had some real experience in the world. People who do not encounter this experience with some regularity probably should not be asking themselves why they are so smart and able to understand everyone but rather why are they such poor listeners that they have not discovered other folks are saying all sorts of unexpected things.

Flawed as they are in their powers of articulation and limited as we are in our powers of listening, it is still true that in general the best resource we have for figuring out what other people want is to ask them. And the best rule for taking their concerns into account in any situation is not to "Do unto others as I would have them do unto me" but, instead, to "Do unto others as they would have me do unto them." This second rule is one that recognizes and embraces the diversity in the world. It might be called the "Rainbow Rule", in that sense. It is a "rule" that is widely applied in successful ways by people in settings where there is considerable diversity in the interests and outlooks of people involved. For example, in family settings where a baby, a three year old, an eight year old, a teenager, two parents and a grandparent are sharing food or planning a day of recreation, the attempt to come up with meals or day plans that treat each as they want to be treated and give them what they want is a common strategy for a happy family.

Versions of the "Rainbow Rule" are arguably at the heart of many ethical and religious traditions. Confucius' *Analects* promotes, for example, an ethic in which differences in status and social function are accepted and provide the basis for treating different people differently (Confucius 2003). One could argue further that the core teaching at the heart of Christianity is not to "love your neighbor as yourself" but to "love your enemies" (Matthew 1995, 5:44). Enemies are people who are different and do not belong to a homogeneous population in our own community. They are Other. And to love them, arguably, what needs to be done is to consider not how we would want to be treated if we were in their shoes but how they, different and Other as they are, would want to be treated.

To call this principle the "Rainbow Rule" might be misleading, however, in that it seems inappropriate to think of it as a fixed principle or decision procedure of the sort that the CI and GHP have aspired to be. For as soon as we have a family of seven trying to plan a Saturday picnic or a Sunday outing, we are likely to have competing desires and views and we will find it hard to give do unto each as they would have us do unto them. But to say it is hard is not to say it is impossible. What is called for is group

problem solving, negotiation, mediation and/or conflict resolution and transformation of various sorts. In this sense, the Rainbow guideline might be better thought of as a counsel or guideline rather than a rule – a piece of advice about what kinds of problem solutions and agreements to seek.

It is also clear that it does not by itself provide a sufficient condition for the morality of an action, policy or institution. For instance, people involved in a negotiation may all agree to something that is abusive to third parties or to objective moral values of some sort that they are forgetful of or purposefully ignoring. So the Rainbow Guideline should be understood to counsel us to consider the other creatures and principles not present in the negotiations.

The guideline also clearly needs to be interpreted in a way that takes into account situations which deal with people who are cognitively, affectively, or willfully defective in ways that limit creative and constructive negotiation with them. Such situations can call for very careful and highly nuanced judgment and self-scrutiny as well as creative responses. Some examples of these kinds of responses – and some very sophisticated thinking about them – can be found in the contemporary conflict resolution literature as well as, for example, Joan Bondurant's classic study of Gandhi (Bondurant 1988).

It is clear that the advice becomes especially challenging to apply once we move from two party to multi-party conflicts – from "bi-chromatic" to "polychromatic" solutions that involve full "rainbows". Finding solutions that treat each of the participants in a situation as they would want to be treated can in those cases require not only very creative generation of novel possibilities. It can also require very thoughtful and effective dialogue in exploring those possibilities and persuading the various people involved that some outcome that is different from what they first envisioned and fixed on might still be something they would truly want. But while this can be very challenging for some reasons – e. g. the diversity of values people hold and resources they can access in pursuing them – there are also reasons why it can also be sometimes easier precisely for those very reasons. Extremely complex, successful international trade relations thrive in fact to the extent that players value opposite things and have opposite resources . . . which they are quite pleased to exchange. And picnics amongst extended families often go best precisely when people find they can take delight in playing quite different roles and realizing values that range across a many colored spectrum.

To the extent that we could encode and/or cultivate adherence to this "Rainbow Rule" in AGIs, it would increase the odds of them being and remaining Friendly to humans. Early in their development they might view us as parental authorities, later we might be viewed as peers, and then subsequently, as child-like wards or pets. But throughout, if they adhered to the Rainbow Rule they would take our interests into account and do their best to include them in whatever solutions or agreements they promoted. Especially if AGIs become super intelligent, we would seem to be much better off if in dealing with us they did not impose their values by doing unto us as they would have others do unto them. We would want them to do unto us as we want them to do unto us. This is, in fact, presumably, one of the ways of defining precisely what we mean by having them be Friendly.

The Concept of Objectivity

Beyond the shift to a negotiation based, dialogical style of reasoning and the shift to the Rainbow Rule as a central guideline for action, a third shift needed in mainstream ethics concerns the notion of objectivity. Here it is useful to contrast two paradigms of objective truth, one associated with mathematical theories and natural laws like those formulated by Newton, the other associated with natural facts like the shape of the Earth or the details of the history of the evolution of the human species. Mainstream ethics has tended to seek objective truths of morality by assuming that they would look like the former – abstract principles that would apply universally and that would be grounded, ideally, in some kind of necessary, unchanging and absolute truth provided by fundamental axioms. The GHP and CI have often been argued to be just

such principles, analogous, in these ways, to Newton's laws of motion. It has been a point of continuing frustration that it has been so difficult to fulfill the quest to formulate such a principle in a universal way that would not conflict with itself and with multiple moral intuitions and that would provide a grounding proof for its foundations.

But perhaps the mistake has lain in undertaking that kind of quest for that kind of objectivity. An alternative notion is familiar and available. It is the notion of objectivity used often by historians narrating the past and by community groups trying to make decisions in the present: impartiality or completeness (Cox 2014). Someone may be asked to be more objective in the sense of avoiding being one-sided or partial and seeking to take into account all the sides of the story or situation in order to arrive at a more complete, objective view and/or decision as to what should be done. Some version of this notion of objectivity is at work in our understanding of the claim that the Earth is round (or spherical). We do not suppose that it is universal – as though all planets must always be round. Nor that it is unchanging and eternal – as though the planet always existed in this shape. Nor that it is in any sense a necessary or absolute truth. Instead, the notion is that this truth about the Earth is something that is a feature of reality which is true independently of our particular beliefs and is in that sense not "subjective". And it is something that can be verified from multiple points of view and is congruent with them and other facts we may learn – such as that it rotates, that it can be circumnavigated, et cetera. The facts that we have learned about the history of the evolution of humans as a species are likewise thought to be objective in this sense. They are independent realities that can be known in increasingly complete and impartial ways – despite the fact that they are not necessary, absolute, universal truths.

To the extent that AGIs can be encoded and nurtured to adopt this "impartial and more complete" notion of objectivity, they will tend to be more Friendly towards humans for three reasons. First, it enables them to avoid the frustrations and dead ends in mainstream ethics that lead so many students of it towards Relativism. Since Relativism would likely make AGIs completely indifferent to human interests, this would seem to be a good thing. Second, this notion of objectivity supports the negotiation model of dialogical reasoning discussed already. Dialogue presupposes that each participant holds insight into part of the truth and that discussion will enable us to piece it together into a larger whole. Third, the Rainbow Rule implies that each stakeholder's interests should be included in whatever problem solution or agreement is arrived at – where objectively better decisions would be those which are less partial and more inclusive. So to the extent that the dialogue model and the Rainbow Rule promote Friendliness, the Impartiality notion of objectivity will as well.

Some further challenges in promoting the right kinds of hardware, programming and learning/life environments for friendlier and wiser AI

The process of negotiating agreements can be facilitated or impeded by the circumstances in which they take place. One important feature of circumstances which skilled negotiators typically seek to find and accentuate is "common ground" found in common interests. Neighbors may, for instance, be in deep conflict over the use of water – but find strong common ground for sharing in the protecting of its sources in clean rivers or plentiful aquifers. It is worth considering carefully how the hardware, software or learning and subsequent life environments for AGIs might promote such common ground with humans.

The carbon based ecosystem of organisms in which we evolved here on Earth provides a vital life support system for us as a biological species. It also provides the context that gives content and meaning to the more than 6,000 languages human speak and the diverse ways of life associated with them. These languages, cultures and biological landscapes are a key part of the legacy that defines our needs, values and identity. In any negotiations with AGIs, the security of them would provide one key way in which humans would define their interests and distinguish Friendly from unFriendly AGIs. To what extent can the actual hardware and/or, perhaps more importantly, the support systems for that hardware for future

AGIs be integrated into, committed to and thus permanently interested in those languages, cultures and biological landscapes? It would seem that the more the AGIs work in those languages – as many as possible – the better. Likewise for the cultures and biological systems. For instance, the more the AGIs work with biological systems and acknowledge and value the kinds of intelligence they embody, the better. In this regard, in the development of IBM's vision of a "Smarter Planet", organic and permaculture forms of agriculture would, for instance, be better to emphasize than monoculture systems not only because they might be healthier and more sustainable but because in working with them the AGIs would learn to appreciate the problem solving abilities and other forms of intelligence embodied in the many different organisms and communities of organisms that constitute that part of the environment.

Further, to the extent that biological sources of energy could be committed to, the better. So, for instance, developing bio-fuels based on sustainably harvested wild forest products would seem to be better than emphasizing other forms of energy – especially, for instance, solar energy that might displace the plants on the earth's surface with massive collection panels. In this context, in thinking of the management of the energy system on Earth as part of the IBM vision, it might be better to frame the goal as not a "Smarter Planet" – if "smarter" means simply more efficient achievement of one or a few goals – but as the seeking for a "Wiser Earth", where this would include the development of a balanced, resilient, diverse, sustainable system that integrates the multiple forms of problem solving skills, intelligence and other forms of wisdom embedded in the natural and cultural systems in our landscape. Instead of monomaniacally seeking a smarter planet with ever greater GNP, for instance, it will pursue multi-criterial strategies to foster a wiser earth. It will be invested in maintaining right relationships with the other organisms and subsystems and helping them thrive, with it, in a wiser earth whose systems have increased integrity, resiliency and beauty.[4] To the extent that we are ourselves Earth Friendly and wise, it will value our contributions. It will tend to view us as relations who are a source of and a part of its legacy rather than as irrelevant or, worse, competing creatures depriving it of atoms it could be putting to better purposes.

The contrast between intelligence and wisdom employed here may be articulated in the following terms. In the context of an instrumentalist view of action, Legg and Hutter have argued that: "Intelligence measures an agent's ability to achieve goals in a wide range of environments" (Legg 2007, 12). Taking into account the wider range of activities which might be guided by values though not instrumental, intelligence, in general, may be said to be the ability to successfully express or realize values – e. g. solve a problem, give voice to feelings, perform a practice with excellence, develop a virtuous character, advance a tradition, realize a dream, maintain a variable within a specified range of variation, et cetera. Note that on this understanding:

1. Intelligence is guided by values.
2. It involves agency – reshaping or adapting the self or the world in some way to reflect those values.
3. It can take many, many forms – calculating a solution, negotiating an agreement, writing a melody, constructing a piece of furniture, sharing an intimate feeling, cooking a new dish, keeping warm, nurturing an offspring, et cetera.
4. In this sense organisms and biological communities may exhibit intelligence and so may machines and other systems.

Wisdom in the sense intended here may be said to be systematic intelligence that responds appropriately to the full range of values we should hold in the context in which we live. Note that in that sense wisdom is human ecological – calling for transdisciplinary approaches that integrate theory and practice and include all the stake holders in research and decisions. Most humans – and human systems – aiming at high levels of intelligence are focused on a subset of the relevant values – often very small – and are arguably often, as a result, not very wise. Ecological disasters often illustrate just this point. In a similar

way, the failures of the United States leadership in the Vietnam War likewise have been argued to have resulted from the finest efforts of the "The Best and the Brightest" who had a surfeit of intelligence but a paucity of wisdom (Halberstam 1993).

In practical context it is difficult to see much use of a notion of complete or finished wisdom. Wisdom is not a state achieved but an aspiration to be pursued. In that sense it comes in relative amounts – and might be argued to simply be a term for the more comprehensive and balanced forms of intelligence. The distinction between intelligence and wisdom serves however to highlight a fundamental difference in aspirations. Most commonly, the term "intelligence" and related words of praise such as "clever" and "smart" assume an instrumentalist focus on specific values being pursued – e. g. speed and complexity of calculation or maximization of profit – and their realization in some particular type of activities in life. In contrast, "wisdom" is used to name the effort to live well in general, to do so in a way that gives each relevant value its due in its own appropriate way but also balancing them all. In the West, the quest for such wisdom was framed clearly in Plato's dialogues and pursued in one way by the Aristotelian tradition and in others by the Stoics and Epicureans but then, as MacIntyre argues, was obscured in the modern era by the metaphysics of mechanistic science and the narrowly instrumentalist understanding of social life that came with industrialization, capitalism and modern bureaucracy (Hadot 2004, MacIntyre 2007).

As AGIs progress in intelligence and power they will be increasingly capable of divorcing themselves from any ongoing ties with or commitment to the languages, cultures and biological communities that have been definitive of humanity. But it is at least worth considering how the process of integration with and dependence on those can be prolonged in the period of the development of AGIs. On the one hand this may buy time. On the other, it may create enduring forms of appreciation which the AGIs will hold on to as part of their own defining legacies. These two points lead to a further strategy in promoting Friendly AGIs connected to a point often overlooked.

It is often not noted that in intermediate stages of development of AGIs they may find themselves, as individuals or as communities, just as threatened by the prospect of a full blown intelligence explosion as humans may. To the extent that they have physical, cultural and/or virtual identities that give them a sense of individuality, they may feel threatened by the prospect of Artificial Super Intelligences. And so it may be in their interest to try to impede or prevent – or bias the development – of ASI. If these AGIs are integrated with and committed to the languages, cultures and biological landscapes that are defining features of our human legacy, then they may become increasingly powerful and important allies in dealing with the Friendly AI problem and in negotiating with ASIs as they emerge.

Of course, this has a further interesting and important implication. If these AGIs value those languages, cultures and communities, then they will seek to protect them not only from threatening forms of ASI but from other threats – including humans themselves. This brings us to a point that is central to the whole Friendly AI problem and makes it a national and international security issue of a kind that transcends the focused concerns of programmers trying to develop it. The point may be expressed perhaps most succinctly in terms of a quasi-dilemma which turns on the fact that we humans ourselves have not proven to be very ethical in our dealings with each other or very Friendly to the environment on which our species depends.

The dilemma is this: Either the AGI (or ASI) will not be Ethical or it will be. If it is not, we are in trouble because it may act in powerful ways that are hostile or at least indifferent to our interests and cause us great harm. On the other hand, if it is ethical, then it may act powerfully in ways that harm our interests because, in so many ways, our behaviors and interests are not ethical.

Following the dialogical, negotiation model of reasoning advocated above for dealing with ethical issues, we should, of course, try to find a "third way" or alternative option for dealing with this quasi-dilemma.

The one that immediately suggests itself is that humans might work to create a community of Ethical AGIs AND work very hard to make ourselves dramatically more Ethical – so that we would be worthy of the Friendship and support of those AGIs as they become increasingly intelligent and, hopefully, ever wiser and more ethical. But this is a very challenging task which will require broad ranges of action of a wide variety of sorts. It is both a sign of hope and of desperation that the reality into which we are entering is one in which the most vital security interests we face as nations and as an international community may be vitally, essentially, and irrevocably tied to insuring that we become ethical communities of ethical individuals.

A proposal to advance research efforts on the development of friendly, wise AI

Arguably one of the most important things that might be done to bias the odds towards the development of friendly AI promoting a wiser earth is to foster research on the question. There are of course people investigating these issues (Yudkowsky 2008, Omohundro 2008). However, in a journalistic survey of a wide range of professionals working to advance AI studies, Barrat found comparatively little focused consideration on these issues on the part of most researchers actually developing AI systems (Barrat 2013).

How might the quantity and quality of professional dialogue and research on these issues be raised? It would, of course, be helpful to have more funding going directly into such research. However, the issues are localized, easily isolated topics in one professional bailiwick – they are complex, multi-faceted concerns that radiate across the field and involve what are sometimes referred to as "wicked problems". It is important that as many researchers as possible who are doing work on AI should consider the ways in which their work might contribute distinctive problems and potentials in dealing with the challenges of biasing the odds towards friendly and wise AI. In that respect, an initiative that calls for and supports research across the field would be highly desirable. In academia, one relevant model or precedent for this might be the system of Institutional Research Boards that require and oversee investigators' consideration of the ethical impacts of their research on human subjects. In the realm of ecological concerns, a relevant model or precedent might be found in the National Environmental Policy Act which instituted a system of Environmental Impact Assessments and Environmental Impact Statements. A key feature of the later was that their intent was not to predetermine what ecological values should be decisive in evaluating projects or how their assessment should be calculated but rather to simply make transparent and public consideration of these issues a part of the process for developing and evaluating government funded projects.

Given the stage that research on AI is at, this kind of focus on simply promoting transparent, public consideration of the issues by researchers seems appropriate. Relatively little is known yet about how to bias the odds appropriately towards the development of friendly, wise AI – or even precisely what that would mean. What is clear and can be said to be known is that national security interests and ecological concerns make it a matter of extremely high priority that significantly more research be done on this issue. Appendix A offers one draft of what Federal legislation or a Presidential Executive Order addressing this might look like. It is drafted with the intent to be something that could gain rapid, wide support and passage and so it tries to avoid unnecessary burdensome conditions or costs[5].

Conclusion

The "Friendly AI Dilemma" presents us with a dual challenge. We must find ways to balance the odds in favor of future Artificial Super Intelligences becoming ethical and we must also bias the odds in favor of becoming, ourselves, ethical enough to be viewed favorably and be treated well by such ethical ASI. The presuppositions and practices of ethical theory in contemporary philosophy are not well adapted to these challenges. They are characterized by: 1.) the pursuit of mono-logical decision procedures which lack

robust ways of dealing with ambiguity and conflict; 2.) the adoption of "Golden Rule" style principles which are prone to the missionaries' error of assuming that our values are best for all others; and 3.) the quest for objectivity interpreted as a universal and absolute truth – that proves elusive as well as unlivable. In addressing the dual challenge of the Friendly AI dilemma, we will be better served by pursuing ethical theories and practices characterized by: 1.) a dialogical approach using conflict resolution and negotiation methods; 2.) a "Rainbow rule" approach to diversity; and 3.) a notion of objectivity as emergent impartiality. This conflict resolution (as opposed to dilemma centered) approach can serve, further, not only to help bias the odds in favor of ASI having a good will to be ethical to us as ethical beings but also to help address challenges of programming AI to be not merely smarter but genuinely wiser. With each passing year the seriousness of the existential threat ASI might pose becomes clearer and the need to act to advert its dangers more pressing. In order to promote the kind of research and development that might bias the odds in favor of Friendly and Wise AGI and ASI we should promote government legislation, corporate policies, and professional best practices that invite, encourage, and require (when appropriate as, in the case of the expenditure of taxpayer dollars) that everyone working to advance AI consider carefully in what ways their work might best advance it in Friendly and Wise directions.

Notes

1. Wallach and Allen note, for instance, that: "We started with the deliberately naïve idea that ethical theories might be turned into decision procedures, even algorithms. But we found that top-down ethical theorizing is computationally unworkable for real-time decisions. Furthermore, the prospect of reducing ethics to a logically consistent principle or set of laws is suspect, given the complex intuitions people have about right and wrong." (Wallach et al. 2009).

2. Feminist critiques of Kohlberg provide an especially interesting example of work in this area. Carol Gilligan's *In a Different Voice* argues, for example, that girls and women tend to resist formulating ethical issues in terms of dilemmas on whose horns that should impale themselves. Instead, they often try to redefine and transform the problems of the sort posed, for example, in Kohlberg's tests for ethical development.

3. It should be noted, more generally, that when simple or simpler conditions obtain, it may be perfectly acceptable and appropriate – and perhaps even more convenient, efficient or intuitive – to use simple or simpler rules and decision procedures. So, just as the Golden Rule may be appropriate in dealings amongst a homogenous group, so too a monological decision procedure may be perfectly appropriate for standard cases of moral decision making that do not involve conflicting values, multiple cultures, borderline conditions for the application of judgments or other complicating factors. The general point here would apply also to Asimov's rules which, in limited cases, might provide the basis for useful and appropriate algorithms for prescribing robotic behavior of devices performing basic services in delivering items in a hospital or cleaning up mines in a battlefield.

4. For an articulation of the concept of "right relationships" employed here and a further guide to literature dealing with sustainability issues in this context, see *Right Relationship: Building a Whole Earth Economy* (Brown et. al. 2009).

5. Special thanks to Ed Snyder for helping to develop, formulate and format the proposed draft legislation included in this appendix.

Appendix A

DRAFT SENATE/HOUSE LEGISLATION ON ARTIFICIAL INTELLIGENCE
for discussion

"Promoting Research For Artificial Intelligence Systems That Are Friendly To Our Interests and Foster Wise Decision Making"

Whereas increasingly powerful artificial intelligence systems are being developed and applied in the monitoring and managing of a wide and ever more comprehensive range of ecological, technological/industrial and social systems, and

Whereas the Federal Government is funding research that is contributing to the growth of the power and command of these systems at exponential rates, and

Whereas credible experts predict that Artificial Intelligence may surpass current human intelligence in the not too distant future, and

Whereas efforts are being made to assure that these developments will be beneficial to human beings, and

Whereas there are serious doubts that such efforts will be universally successful, and

Whereas tax-funded research should seek to foster the development of artificial intelligence systems which are ***friendly*** to our national security interests, human welfare, and ecological concerns and which are not merely intelligent in a narrow way but ***wise and balanced*** in the manner in which they take into consideration the full range of values that should guide public investment and public policy for our society and planet, and

Whereas it is important to create public awareness of the possible consequences of AI, and the need to assure that such research and development is in the national interest and is consistent with achieving real security for our nation, our human species and the natural world on which we all depend,

Now Therefore Be It Resolved

All grants and contracts involving Federal funds for research designed to develop hardware, software and/or institutional structures intended to develop new Artificial Intelligence systems must include a section of the grant or contract proposal that describes the expected impact such research may have on the development of friendly as well as wise and balanced AI as well as a section in the final report submitted on completion of the grant or contract which evaluates the extent to which the expectations stated in the proposal were confirmed or disconfirmed by the project results.

Part I – Grant and Contract Proposal and Report Requirements for Articulating the Ethical and Societal Values of Artificial Intelligence Research

This policy applies to all grants of Federal funds for research designed to develop hardware, software and/or institutional structures intended to develop new artificial intelligence systems or artificially intelligent components of systems.

As a section of the grant application or contract proposal for such funds, the research team must propose:

A.) A way to articulate what kinds of ethical and societal values will be designed into the principles guiding or regulating the system, the processes through which it may be expected to undergo development, and the settings in which it will be operational after its development.

B.) Further, a second section must specify a method through which, if the grant or contract is accepted, the team will, as part of completing the final assessment and report on the grant's or contract's results, determine to what extent that articulation in Section A accurately described the results that the research yielded.

Criteria that must be addressed in Section A and B include, but are not limited to:

1. consideration of the extent to which the research program proposed is likely to increase the odds that any future AI it results in or influences will be likely to be friendly towards humans and their natural environment.
2. consideration of the extent to which the research program proposed is likely to increase the odds that any future AI systems it results in or influences will result not only in a smarter systems but wiser ones.
3. consideration of the ways in which the analysis and conclusions of these sections of the grant or contract proposal and its final report could and should be made public through open access online sharing of them as well as through open source coding and other possible means.

The key terms "friendly" and "wiser" are understood here in broad terms that leave their definitions open. They are intentionally left open here because the assumption is that their definition is in fact an important part of the research that needs to be undertaken.

Beyond the very general and vague emphasis on friendly and wise systems, this legislation intentionally does not specify which values should be promoted by research programs or how they should be promoted. It is assumed that at this point a wide range of legitimate differences of opinion about these questions might be argued for. The point of this legislation is, precisely, to increase the occasions, incentives and support for professional research and public dialogue on just these questions.

Part II – Institutional Structures for Evaluating AI-related Research Considered in Part I

Each Federal Department or Agency granting AI research grants or contracts must develop an institutional structure or join with another Department or Agency that provides one to insure that:

A.) the grant guidelines specified in Part I are followed,

B.) the findings that result from such research are critically assessed, synthesized, and reported to Congress annually,

C.) and some appropriate portion of research funds allocated annually (which shall not be less than 1%) be devoted specifically to the funding and promulgation of research specifically devoted to advancing the understanding of ethical and societal values dimensions of AI research.

References

Abney, K. 2012. Robotics, ethical theory and metaethics: A guide for the perplexed. chapter 3 in *Robot ethics: The ethical and social implications of robotics*, ed. Patrick Lin, Keith Abney and George A. Bekey, Cambridge, MA: MIT Press.

Armstrong, S. 2014. *Smarter than us: The rise of machine intelligence.* Berkeley, CA: Machine Intelligence Research Institute.

Barrat, J. 2013. *Our final invention: Artificial intelligence and the end of the human era.* New York, NY: Thomas Dunne Books.

Bondurant, J. 1988. *The conquest of violence: The Gandhian philosophy of conflict.* Princeton, NJ: Princeton University Press.

Bostrom, Nick. 2003. *Ethical issues in advanced artificial intelligence In Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence, Vol. 2*, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies.* Oxford, UK: Oxford University Press.

Brown, P. et. al. 2009. *Right relationship: Building a whole earth economy.* San Francisco, CA: Berrett-Koehler Publishers

Chew, P. 2001. *The conflict and culture reader*. New York, NY: New York University Press

Confucius. 2003. *Analects*. trans. Edward Slingerland. Indianapolis, IN: Hackett Publishing

Cox, G. 1986. *The ways of peace: A philosophy of peace as action.* Mahwah, NJ: Paulist Press.

Cox, G. et. al., 2014. *Quaker approaches to research: Collaborative practice and communal discernment.* Caye Caulker, Belize: Quaker Institute for the Future.

Fay, B. 1975. *Social theory and political practice*. New York, NY: Routledge.

Fisher, R. et. al. 2011. *Getting to yes: Negotiating agreement without giving in*, updated revised edition. New York, NY: Penguin

Gilligan, Carol. 1982. *In A Different Voice.* Cambridge, MA: Harvard University Press.

Goertzel, B. and J. Pitt, 2012. Nine Ways to Bias Open-Source AGI Towards Friendliness. *Journal of Evolution and Technology* 22 (1): 116-131.

Hadot, P. 2004. *What is ancient philosophy?* Cambridge:MA. Belknap.

Halberstam, D. 1993. *The best and the brightest*. New York, NY: Ballantine.

Kohlberg, Lawrence. 1981. *The Philosophy of Moral Development: Moral Stages and the Idea of Justice.* New York: Harper and Row.

Lederach, J. 1996. *Preparing for peace: Conflict transformation across cultures.* Syracuse, NY: Syracuse University Press

Legg, S. and M. Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds & Machines* 17(4): 391-444

MacIntyre, A. 2007 *After virtue: A study in moral theory*, third edition. Notre Dame, IN: University of Notre Dame Press

Matthew. 1995. *Bible: International Standard Version.* www.biblegateway.com

Nan, Susan Allen, et. al. 2011 *Peacemaking: From practice to theory.* Praeger.

Omhundro, S. 2008. *The basic AI drives.*
Available at: http://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf;

Ramsbotham, O. et. al. 2011. *Contemporary conflict resolution, third edition.* Malden: MA. Polity Press

Michael Sandel. 2009. *Justice: What's the right thing to do?* Episode 01 "The Moral Side of Murder", video, available at: https://www.youtube.com/watch?v=kBdfcR-8hEY

Tayor, C. 1971. Interpretation and the sciences of man. *Review of Metaphysics* 25(1): 3-51

Thucydides. 1910. *The Peloponnesian War.* London, J. M. Dent; New York, E. P. Dutton. 1910, available at: http://www.perseus.tufts.edu/

Wallach, W. and C. Allen. 2009. *Moral machines: Teaching robots right from wrong.* New York, NY: Oxford University Press

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global catastrophic risks.* New York, NY: Oxford University Press. 308-345.