Engineering Better Ethics into Human and Artificial Cognitive Systems

John Celona(✉)

Decision Analysis Associates, LLC, San Carlos, CA, USA
JCelona@DecisionAA.com

Abstract. Evolutionary ethics presents a hypothesis for understanding ethics as an evolved social behavior which develops according to the advantages or disadvantages it creates for an individual in its environment. Using this framework, designers of human or artificial cognitive systems can identify the ethical standards they wish to follow and their system to promote, and design into the system the legal, social, or prudential incentives or disincentives to promote the desired choice of ethical standards.

Keywords: Ethics · Evolutionary ethics · Human system design
Artificial cognitive system design

1 Introduction

As I discuss in a chapter written for a forthcoming book [1], designers and engineers of human and artificial systems grapple Frankenstein-like with the ethical implications of their systems and their use. Adding a cognitive dimension (where the system evaluates and possibly learns) furthers the complexity and difficulty of considering consequences and their ethical implications, both anticipated and unanticipated. These consequences may be realized in the operations of the technology itself, or in the system's effect on the behavior of people interacting with the system.

Designers and engineers striving to meet these challenges face daunting challenges in identifying, abstracting and applying the relevant ethical rules. Many lifetimes can be spent studying the various theories and approaches to ethics.

Even then, it is not obvious how to apply those learnings to a human system (like the law or a government benefit) or an artificial learning system. Examples of the latter in a broad sense include Facebook and Google algorithms, facial recognition, cyber-security, and deep learning systems to dispatch human safety inspectors. What are the ethical implications of a boiler in a school exploding after a deep learning system indicated that a safety inspection was not required?

To meet these challenges, we propose a straightforward approach: a hypothesis useful both to define the desired ethical standards and to predict whether the system results (whether from action of the human or technology components) will meet those standards. This approach can be used to engineer the desired ethical standards into the system. The hypothesis is called *Evolutionary Ethics*.

W. Karwowski and T. Ahram (Eds.): IHSI 2019, AISC 903, pp. 56–62, 2019.
https://doi.org/10.1007/978-3-030-11051-2_9

2 Evolutionary Ethics

The hypothesis is straightforward: what if ethics were an evolved social behavior? Like all hypotheses, the question of whether this is true or false does not apply: a hypothesis can only be disproven, never proven. Rather, the relevant question is whether the hypothesis makes testable predictions for observable phenomena of interest. If so, we use it until a better one comes along.

If ethics were an evolved social behavior then, like all evolved characteristics (whether physical or behavioral), behaviors which created advantages for a species in its environment would be passed along either genetically or through teaching. Those which disadvantaged the species would be weeded out.

There is considerable evidence that ethics exists in other species [2] and, indeed, Darwin himself thought that ethics followed the same evolutionary process as other physical characteristics and behaviors. "The following proposition seems to me in a high degree probable- namely, that any animal whatever, endowed with well-marked social instincts, [citation omitted] the parental and filial affections being here included, would inevitably acquire a moral sense or conscience as soon as its intellectual powers had become as well, or nearly as well developed, as in man." [3]

Subsequent work reviewed and conducted by Bekoff and Pierce [2] found an ethical and moral sense in a variety of social species, including elephants, wolves, dogs, rats, cats of all sizes, bats, monkeys, etc. They organize their work on the moral lives of animals into "the *cooperation* cluster (including altruism, reciprocity, honesty, and trust), the *empathy* cluster (including sympathy, compassion, grief, and consolation), and the *justice* cluster (including sharing, equity, fair play, and forgiveness)." [4]

Some of the advantages conferred by cooperation, empathy, or justice are straightforward, others less so. Sharing food may confer an immediate disadvantage (you get less to eat), but a long-term advantage in prompting others to share with you in the future when you may be of need. Many ethically-implicated behaviors share this pattern: they confer an immediate disadvantage to the individual at the promise of a future benefit when others reciprocate. Indeed, this is the essence of the "Golden Rule:" "So whatever you wish that men do to you, do so to them…" [5].

Other advantages are less obvious. Where is the advantage in foregoing an immediate opportunity or sacrificing ones' own interests when there is no foreseeable prospect of future reciprocation by others?

One possible advantage is avoiding sanction for unethical behavior and, indeed, many social species in addition to man have sanctions for unethical behavior (e.g., expulsion from the pack.) In human society, possible sanctions may be by the legal system or social (shunning, disapproval, etc.).

However, actions may raise ethical implications even though they are perfectly legal and socially approved. You have the legal right to kill in self-defense (subject to limitations) and it is socially acceptable to do so, but you may feel it is wrong to do so nonetheless. The sixth commandment is "[t]hou shalt not kill," without any reference to self-defense. Many soldiers experience trauma from society's requirement that they fight and kill in war, even those who do so by controlling remote drones thousands of miles away from the fight. [6] Is there a "moral injury" unethical behavior inflicts on

the perpetrator apart from and in addition to the immediate and possible future tangible consequences? Could "moral injury" be part of the evolved response to promote more ethical behavior?

If ethics evolve, we need a framework for specifying ethical standards that (1) captures how the ethical implications of actions may advantage or disadvantage individuals and systems and drive future behavior; (2) clarifies the ethical implications of the behaviors that may result; and (3) enablers designers and engineers to think hard about the ethics they are designing into the system and make informed choices.

3 A Framework for Specifying Ethical Standards

To describe this framework, we need to consider both the objectives of ethics (what do they accomplish?) and for whom (the individual or others). The objectives dimension needs to show how actions may advantage or disadvantage an individual. The "who" dimension needs to capture the self- or other-regarding effects.

For the first dimension (what goal?), consider Maslow's hierarchy of needs. [7] The needs he identified (in order of priority) are: Physiological Needs (food, water, temperature), Safety and Security, Love and Belonging, Self-esteem, and Self-actualization. For the second dimension (for whom?), consider a scope beginning the individual and progressing to larger scope: family, tribe or nation, species, possibly then warm-blooded life and then all life. These two dimensions describe a continuum of possible ethical standards, as shown in Fig. 1.

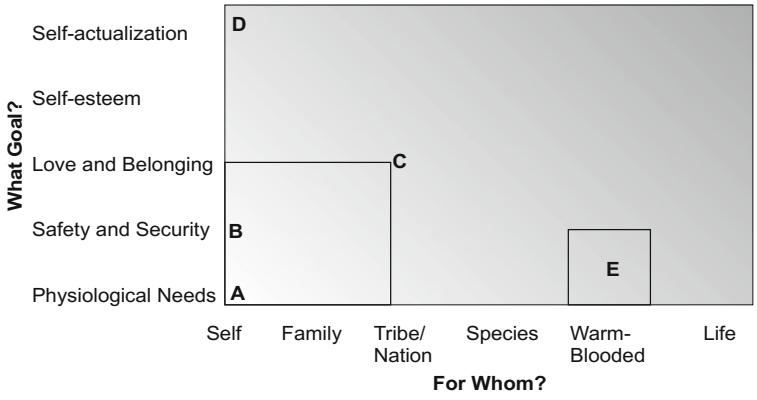

Fig. 1. A continuum for describing possible ethical standards

Given those two dimensions, an individual's ethical standards can be specified by picking a point on the graph. The area between that point and the axes are the interests that person wishes his or her ethical standards to respect. Actions which violate those interests would be unethical.

For example, an individual at point *A* would only be concerned with obtaining food, water, and maintaining temperature. An individual at point *B* would be developed enough to recognize danger and react to it. An individual at point *C* adds other-regarding social behavior to consider the needs of its entire group (family or not). A human sociopath (by definition incapable of feeling for others) would fall at point *D*.

One could also specify unusual ethics. An individual who felt it was unethical to kill animals but fine to kill people would have standards described by the area around point *E*. For the purpose of discussing how system design impacts groups of individuals, we'll focus on the progressive addition of objectives and scope followed by most individuals and groups rather than outliers.

4 Methods of Motivating Observance of Ethical Standards

What drives individuals to follow one ethical standard over another? For action to have an ethical implication, there has to be a choice. You cannot accuse a bacteria which kills its host or a vine which kills the tree it grows on of being unethical because they have no choice; they are only following rules laid down in their genes. Howard and Abbas follow this approach: "**Ethics** are your *personal* standards of right and wrong. Your code of *proper behavior*." [8] An action is ethical if it is right given your choice of ethics. The question, then, is how to drive the choices of ethical standards individuals follow.

One may think of three progressive levels of incentives. Legal systems enforce minimum standards at the risk of formal sanction. Social mores enforce standards at the risk of disapproval or shunning. The most expansive incentive is what an individual perceives to be in his or her best interests, which can be termed prudential behavior. These three methods are illustrated in Fig. 2.

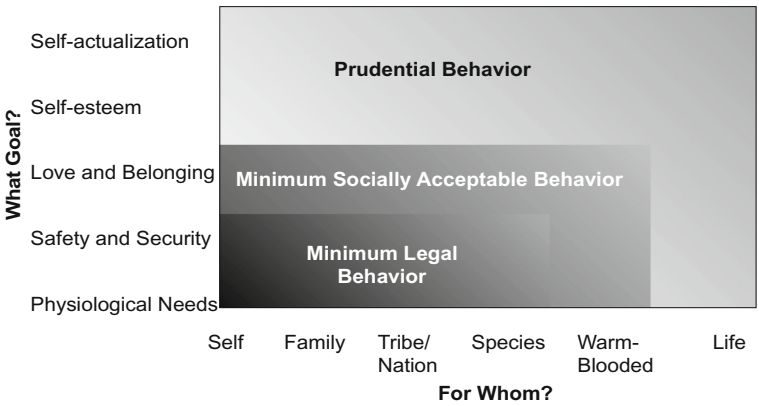

Fig. 2. Progressive methods of incenting choice of ethics

Adhering to one's ethical standards makes one feel good about oneself, helping self-esteem and self-actualization. One the other hand, not living up to them damages both of these, and may inflict the sort of "moral injury" described above. Thus, from this perspective, following one's ethics furthers one's prudential interests even when the behavior deviates from societal or legal standards.

Many difficult ethical dilemmas arise from conflict between acting according to one's lower-level or higher-level goals. Should you lie to get the deal or job? Should you report possibly illegal activity by your best customer, though it may cost you their business? Plus risking future social or legal sanction for your action or inaction.

More serious ethical dilemmas arise with choices sacrificing your own interests to help another, or hurting another's interests to help yourself. Should you donate a kidney to a friend in dire need of a transplant? Should you split your inheritance with disinherited siblings for whom the sum would be life-changing (though it is not for you), even though the deceased specifically excluded them from his will?

At a system level, the question is whether incentives in the system promote and reward higher or lower-level choices of ethics. For example, the rules of professional responsibility require attorneys to zealously pursue any legal objectives their clients may have, not to seek the truth or a just outcome. [1] Likewise, some of mishaps in testing self-driving cars seem to indicate that their operation can lead to greater driver inattention and collisions that most drivers would avoid when driving unassisted. How are system designers to consider the ethical implications in their designs?

5 One-Stage Ethical Design for Human Systems

Considering the ethical framework and incentive structure described above, ethics can be designed into a human system as follows:

1. Using the ethical framework, specify the ethical standards you wish the system to promote.
2. Align the incentives produced by the system (legal, social, and prudential) with the chosen ethical standards.

Legal incentives are a matter of writing the rules and enforcement mechanisms for the desired behaviors. For social incentives, clear identification of undesired behaviors and signaling to others is needed to invoke social sanctions. For prudential incentives, consider whether the system discourages or rewards higher-level ethics (an example of the latter is tax deductions for charitable donations), or unethical behavior that damages higher-level needs (e.g., lying in insurance claims).

6 Two-Stage Ethical Design for Artificial Systems

In designing artificial cognitive systems, a two-stage approach is needed. First, consider the behavior of the artificial portion of the system. Then, consider the impacts of the system on the creatures (including people) which interact with it.

Regarding the system itself, it has no ethics. It simply follows the rules programmed into it. Those rules may prioritize certain feedbacks over others in determining next steps but, barring artificial consciousness (of which there is currently no prospect, a multitude of science fiction writers notwithstanding), the system won't consider whether its actions are right or wrong—only whether they optimize according to the value functions defined for it.

What actions are the system optimized to take? And what impacts do those actions have on the creatures interacting with it? What incentives do those action create for the creatures interacting with it?

Given those incentives, evaluating the impact on people interacting with the system is the same process as described above, just with some of the inputs provided by the actions of an artificial cognitive system in addition to the inputs provided by the relevant human systems and by nature. If this second step identifies incentives for ethically undesirable behaviors, the artificial or human portions of the system may need to be revised to produce different incentives.

7 Summary

Evolutionary ethics presents a framework for understanding ethics as an evolved social behavior which, like all products of evolution, develops according the advantages or disadvantages it creates for individuals in their environment. This framework describes a continuum of possible choices of ethical standards. Ethical standards may be promoted and enforced through legal, social, or prudential means.

To design ethics into human or artificial cognitive systems:

1. If applicable, consider the actions optimized for in the artificial system and their impacts on people interacting with the system.
2. Include these impacts in the total array of effects, incentives, and disincentives, including those from the relevant human systems and from nature.
3. Identify the desired choice of ethical standards.
4. Evaluate the ethical standards promoted by the entire array of incentives and, if necessary redesign the artificial or human portions of the system to incent the desired choice of ethics.

Acknowledgments. The author would like to Dr. Ali E. Abbas of the Neely Center for Ethical Leadership and Decision Making at the University of Southern California for prompting and supporting my thinking and writing on this topic, and Dr. Ronald A. Howard of Stanford University for his insights which spurred my ethical inquiries beginning forty years ago.

References

1. Celona, J.: Evolutionary ethics: a potentially helpful framework in engineering a better society. In: Abbas, A., Gee, S. (eds.) Next Generation Ethics: Engineering a Better Society. Cambridge University Press, Cambridge (2019)

2. Bekoff, M., Pierce, J.: Wild Justice: The Moral Lives of Animals. The University of Chicago Press, Chicago (2009)
3. Darwin, C.: The Descent of Man, and Selection in Relation to Sex, 2nd edn, p. 57. J. Murray, London (1874)
4. Bekoff, M., Pierce, J.: Wild Justice: The Moral Lives of Animals, p. xiv. The University of Chicago Press, Chicago (2009)
5. Throckmorton, B.H.: Gospel Parallels: A Synopsis of the First Three Gospels, 4th edn, p. 29. Thomas Nelson, Nashville (1979). (Citing Matt. 7:12 and Luke 6:31)
6. https://www.nytimes.com/2018/06/13/magazine/veterans-ptsd-drone-warrior-wounds.html
7. Maslow, A.H.: A theory of human motivation. Psychol. Rev. **50**(4), 370–396 (1943)
8. Howard, R.A., Abbas, A.E.: Foundations of Decision Analysis, p. 781. Pearson Education Inc., New York (2016)