

Practical Benevolence – a Rational Philosophy of Morality

Stefan PERNAR ^{a,1}

^a *Director Corporate Security Office, Siemens Ltd., China*

Abstract. These arguments develop the foundation necessary for realizing the logical maxim in Kant's categorical imperative[12] based on the implied goal of evolution[22]. On that basis it is demonstrated that moral behavior is an emergent phenomenon among evolving interacting goal driven agents. In conclusion interaction principles are derived that can serve as basis for a moral philosophy based in rationality.

Keywords. benevolence, emergence, evolution, game theory, morality, rationality

0. Introduction

Questions of morality; how individuals perceive themselves living and expect others to live their lives, have been pondered as well as fiercely debated by philosophers for millennia and differences in opinion have resulted in countless conflicts and unimaginable suffering.

Answering the question of what one should do by establishing morality as a rigorous science grounded in reality thus holds the potential for greatly improving the human condition.

1. Evolution

Evolution is the gradual accumulation of complexity by chance mutation and non chance retention[4] of self replicating units of information. On the chemical level the units of information are represented by genetic code in the form of DNA encoded by four distinct base pairs[3] active in protein synthesis. On the cognitive level the units of

¹ Corresponding Author: Stefan Pernar, 3-E-101, Silver Maple Garden, Cai Hong Lu #6, 100015 Beijing, China; Email: Stefan.Pernar@gmail.com

information are being represented as neural patterns of memetic[17] code in the form of varying firing thresholds in the synaptic junctions interconnecting neurons and are active in information processing of environmental stimuli guiding an individual's interaction with it through cognition[13]. On the genetic level replication takes place through reproduction while on the cognitive level replication takes place through conformist as well as payoff-biased transmission[8].

Following the mechanism of natural selection those information carrying units that increase an individual's inclusive fitness[7] based on the effects they have on the individual's interaction with its environment are being reinforced by having a positively correlated contribution to an individual's ability to pass on said information carrying units. Those information carrying units that do not contribute as much to an individual's ability to pass on said information carrying units or are detrimental to it will eventually go extinct.

1.1. Metasystem Transition Theory

A metasystem transition is the emergence, through evolution, of a higher level of control[2]. A metasystem is formed by the integration of a number of initially independent components, such as molecules, cells or individuals, and the emergence of a system steering or controlling their interactions. As such, the collective of components becomes a new, goal-directed entity, capable of acting in a coordinated way. This metasystem is more complex, more intelligent, and more flexible in its actions than the initial component systems.

1.2. The evolution of cognition in the context of metasystem transition theory

The evolution of cognition can be understood as a series of metasystem transitions[22] with each transition resulting in a crucial boost to an individual's potential for further increasing its inclusive fitness by evolution. In the current model the following metasystem transitions can be identified:

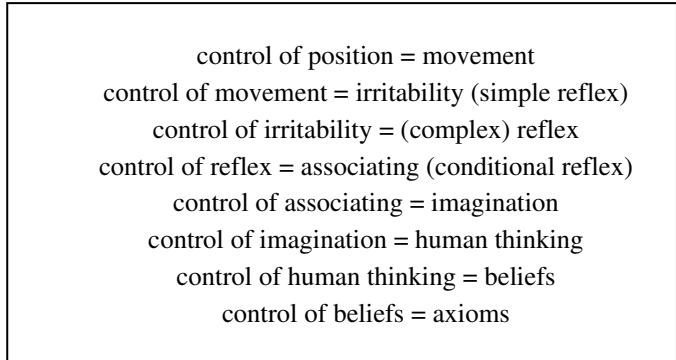


Figure 0. Metasystem transitions in the evolution of cognition

1.3. Beliefs as fitness indicators

When examining the various beliefs held by individuals, different cultures around the world and over the centuries, it becomes clear that belief content is very diverse and can potentially be anything[5]. The fact that each of us holds a particular set of beliefs seems like a mere coincidence – a set of chance mutations acted upon by evolution - since would we have been born in a different time, a different culture, or had different experiences, we would have adopted a set of beliefs in accordance with the axioms of the time.

This suggests that human beings are hard coded genetically with the ability to hold an essentially random set of beliefs. Actual belief content however is arbitrary and thus becomes a fitness indicator for natural selection to weed out unfit beliefs and allow for the evolution of ever fitter belief systems by the mechanisms of cognitive evolution as described above. These belief systems not only determine an individual's fitness but the fitness of the group sharing said belief system that the individual belongs to[9, 10].

All subjective notions of what is good or bad, a virtue or a vice as well as right or wrong thus becomes that set of subjective beliefs that merely did not go extinct yet. Beyond a small set of narrow beliefs however, this insight does not help determining what actually is fitter than another belief, since as long as natural selection has not taken its pick, all as yet not extinct beliefs have a chance to out-exist the others.

1.4. Unjustified mechanism for belief control

Due to the relatively high increase in inclusive fitness for individuals that show the capacity of successfully manipulating other individual's axioms, evolution has caused an arms race between individuals with an ability for manipulation – commonly referred to as individuals possessing charisma - and individuals who can successfully resist such

manipulation resulting in an explosion of the human capacity for rational thought as detection mechanism for manipulation without grounding in reality apparent in our human ancestor's increase in brain size over the past few million years[2]. The end result being the human species - homo sapiens sapiens – the species of the 'double wise people', us.

1.5. Justifying beliefs through science

The scientific method arose from the desire to justify particular beliefs by assigning a degree of certainty with which a particular opinion is true via the process of formulating an hypotheses, designing repeatable experiments to gather entangled evidence in support of said hypotheses[16] and updating ones beliefs about reality using Bayesian reasoning[1]. This method has been expanded by adding the requirement of falsifiability in the critical rational school[20]. The quality – or beauty - of an hypothesis is thus determined by how hard it is to disprove it rather than how well it can be supported by the gathered evidence.

1.6. Evolution's implicit goal

Evolution does not have an explicit goal but the implicit goal of evolution to increase fitness can be derived from the above arguments[22]. From examining what an increase in fitness actually constitutes, it can be concluded that an increase in fitness is equivalent with an increase in the ability of a unit of information to ensure its continued existence.

One could then form the hypothesis that that is good what increases fitness[18] or put another way that that is good what increases a unit of information's ability to ensure its continued existence.

1.7. Deriving evolution's utility function

To justify said claim scientifically one would have to prove it using critical rational methods. For that purpose let us assume the following implied axiom of the above hypothesis:

existence is preferable over non-existence

Where existence is defined as the ability of being perceived. The above axiom seems like a reasonable assumption as a rational individual disagreeing with this claim would have to consequently work towards its own non-existence or put another way strive for self annihilation from perceivable reality. As a next step the above axiom can

be transformed into an utility function or maxim of a goal driven acting agent as following:

ensure continuous co-existence

In this form the utility function becomes an unobjectionable maxim, as neither the agent having said goal can rationally object to it without removing itself from perceivable reality or striving towards being removed by other agents from perceivable reality. Nor can another agent interacting with an agent having said maxim object to it without either wanting to annihilate itself or striving for annihilation by other agents.

2. Morality as rational obligatory survival strategy

As Kant pointed out, a rational philosophy of morals would have to satisfy the categorical imperative:

“Act only according to that maxim whereby you can at the same time will that it should become a universal law.”[12]

Inserting the previously developed unobjectionable utility function as maxim M! in the categorical imperative it can serve as basis for a rational philosophy or morality.

2.1. Relativistic irrationality

Assume two agents[21] $A(i)$ each one with a utility function[14, 15] $F(i)$, capability level $C(i)$ and no initial knowledge as to the other agent's F and C values. Both agents are given resources and are tasked with devising the most efficient and effective way to maximize their respective utility with said resources.

Scenario 1: Both agents have fairly similar utility functions $F(1) = F(2)$, level of knowledge, cognitive complexity, experience - in short capability $C(1) = C(2)$ - and a high level of mutual trust $T(1 \rightarrow 2) = T(2 \rightarrow 1) = \sim 1$. They will quickly agree on the way forward, pool their resources and execute their joint plan. Both agent's utility will thus be maximized.

Scenario 2: Again we assume $F(1) = F(2)$, however $C(1) > C(2)$ - again $T(1 \rightarrow 2) = T(2 \rightarrow 1) = \sim 1$. The more capable agent will devise a plan, the less capable agent will provide its resources and both will execute the plan together. Both agent's utility will be maximized again.

Scenario 3: $F(1) = F(2)$, $C(1) > C(2)$ but this time $T(1 \rightarrow 2) = 1$ and $T(2 \rightarrow 1) = \sim 0.5$ meaning the less capable agent assumes with a probability of 50% that $A(1)$ is in fact a deceptive self serving optimizer who's difference in plan will turn out to be decremental towards maximizing $A(2)$'s utility function while $A(1)$ is certain that they

in fact share one utility function. The optimal plan devised under scenario 2 will now face opposition by A(2) although it would de facto be in A(2)'s best interest to support it with its resources to maximize F(2). Correspondingly A(1) will see A(2)'s objection as being detrimental to maximizing their shared utility function.

Based on lack of trust and differences in capability each agent thus perceives the other agent's plan as being irrational from their respective points of view.

Under scenario 3, both agents now have a variety of strategies at their disposal:

1. deny pooling of part or all of their resources
2. use resources to sabotage the other agent's plan
3. deceive the other agent in order to effect how the other agent is deploying strategies 1 and 2
4. spend resources to explain the plan to the other agent
5. spend resources on self improvement to understand the other agent's plan better
6. strike a compromise to ensure an optimal level of pooled resources

Strategy 1 is a rational given under scenario 3. Strategies 2 and 3 are risky, particularly as it would cause a further reduction in trust on both sides if this strategy gets deployed and the other party would find out. Strategy 4 seems like the way forward but may not work with large differences in C among the agents. Number 5 is a likely strategy with a fairly high level of trust. Under uncertainty however, the rational strategy would be strategy 6.

Striking a compromise is trust building in repeated encounters, promises less objection in the future and thus a higher total payoff for both agents increasing over time.

2.2. Absolute rationality

Assuming the existence of an arguably optimal path leading to a maximally possible satisfaction of a given utility function deploying any other plan would be relatively irrational. Such a maximally intelligent algorithm exists in the form of Hutter's universal algorithmic agent AIXI[11]. The mayor problem being however that the execution of AIXI requires infinite resources and is thus impractical. Consequentially all decisions always have to be made under resource constrains and thus uncertainty.

As a result every decision will be irrational to that degree that it differs from the unknowable optimal course of action that the AIXI algorithm would produce. Under uncertainty in regards to another agent's utility function and varying levels of capability among the agents all agents will thus always have to adopt their plans and strike a

rational compromise based on the other agent's relativistic irrationality independent of their capabilities in order to minimize the damage caused by the other agents' objections on the one hand as well as maximizing cooperation on the other and thus maximizing their respective utility function overall.

2.3. Resolution of moral paradoxes

Assuming the maxim M! in Kant's categorical imperative being equivalent to 'ensure continued co-existence' one must assume it to be the implicit utility function of every goal driven acting agent. However, M! is neither explicitly nor in its entirety encoded explicitly in every conceivable agent's self replicating units of information. It follows that M! generally diverges from the explicit utility function F(i) in goal driven acting agents and that those agents whose F(i) best approximates M! have the best chance of ensuring the continued existence of the self replicating units of information making up the agents' phenotypes.

F(i) can be best understood as evolved beliefs in regards to what should guide an individual's actions while M! is what can not be acted against without causing the self replicating units of information extinction in the long run and therefore should de facto guide an individual's actions.

Consider the following two philosophers[23]:

Philosopher 1: "You should be selfish, because when people set out to improve society, they meddle in their neighbors' affairs and pass laws and seize control and make everyone unhappy. Take whichever job that pays the most money: the reason the job pays more is that the efficient market thinks it produces more value than its alternatives. Take a job that pays less, and you're second-guessing what the market thinks will benefit society most."

Philosopher 2: "You should be altruistic, because the world is an iterated Prisoner's Dilemma, and the strategy that fares best is Tit for Tat with initial cooperation. People don't *like* jerks. Nice guys really do finish first. Studies show that people who contribute to society and have a sense of meaning in their lives, are happier than people who don't; being selfish will only make you unhappy in the long run."

- Philosopher 1 is promoting altruism on the basis of selfishness
- Philosopher 2 is promoting selfishness on the basis of altruism

It is a contradiction - a paradox. But only in thought – not in reality. Applying our gained insights into the nature of M! one can explain what is actually taking place as

following: both philosophers have intuitively realized an aspect of M! and are merely rationalizing differently as to why to change their respective F(i) to better approximate M!.

The first one by wrongly applying the term selfishness on the fallacy that a higher paid job contributes only to his personal inclusive fitness by gaining more resources while in reality it contributes to all individual's inclusive fitness as he is taking the job that is considered to benefit society the most.

The second one by wrongly applying the term altruistic on the fallacy that her recommendations are detrimental to her inclusive fitness due to loosing resources by being friendly while it actually contributes to all individual's inclusive fitness as it not only benefits her but other individuals as well.

The solution thus becomes that the classically intuitive concepts of altruism and selfishness are not helpful in this context.

An altruist giving up resources in a way that would lead to a reduction in her inclusive fitness would be irrationally acting against M! thus being detrimental to all other individual's as well as herself.

An egoist acting truly selfish would use resources in a way that leads to a reduction in his inclusive fitness thus being detrimental to himself as well as to all other individual's.

It follows that in reality there is neither altruistic nor egoistic behavior - just irrational and rational behavior in regards to maximizing M!. The differences being in how an agent rationalizes its own perceived irrational behavior or how an agent's perceived irrational behavior is rationalized by others.

2.4.Trust as emergent phenomenon among interacting rational agents

Let us assume a reality in which all agents are rational[19]. In such a reality all agents will adopt M! as their explicit utility function F(i). All agents would use the resources at their disposal to maximize said utility function.

Let us assume further that initially there are varying levels of trust among the agents. Trust being defined as the perceived difference between the other agents F(i) in regards to the own F(i). Since rational agents would have to adopt M! as their explicit utility function, the level of trust can be redefined as the degree of certainty in regards to the other agent's rationality R(i). For would the other agents be rational they would satisfy $F(i) = M!$

A rational agent without information on another agent's $R(i)$ would have to initially be agnostic towards that agent's $R(i)$ and all else being equal assign an unknown probability towards it.

Considering that an individual agent's actions are derived from an agent's explicit utility function $F(i)$, its capability $C(i)$ and its degree of rationality $R(i)$ and that an agent's $R(i)$ is positively correlated to how efficient and effective that agent can transform its resources into utility, the results of an agent's actions will provide an observing rational agent perceiving the results of another agent's actions insights into the other agent's $R(i)$.

Further, a rational agent would not be deceptive since by definition it can will that its maxim become a universal law (see categorical imperative above). As a result interacting rational agents would through the mechanisms of Bayesian reasoning assign an ever increasing probability into the other agent's degree of rationality ending up pooling all available resources to maximize $M!$

2.5. Respect for others

Consider a reality of interacting goal driven agents with heterogeneous levels of rationality $R(i)$ and resources available to each agent. All agents will turn their resources into utility with an efficiency that positively correlates with their respective $R(i)$. Based on their interactions with each other they will form beliefs as to how well their respective $F(i)$ differs from the other agents' $F(i)$ and again dependent on their rationality pool resources with agents they perceive of having a similar $F(i)$. The accuracy of said beliefs will again be positively correlated with their $R(i)$ levels.

Based on these conditions and assuming the shared implicit utility function $M!$ the following will be observed in an iterated evolutionary simulation:

Scenario 1: $F(i) \neq M!$ \Rightarrow All less rational agent will turn their resources into utility in such a way that is detrimental towards their inclusive fitness and thereby either through defensive actions from other agents or through self harming actions will either have to evolve their $F(i)$ to more closely approximate $M!$ or consequently go extinct.

Scenario 2: $F(i) \approx M!$ \Rightarrow All more rational agents will turn their resources into utility in such a way that maximizes $M!$ and by not being deceptive gain trust and maximize support in the process. Additionally in order to minimize other agents' objections as well as maximizing other agents' support more rational agents will use resources to generate creative compromises between relativistic irrational interests of the self and others in the interest of both (see strategy 6 above) in supporting and opposing the other agents' $F(i)$ and thus minimizing said agents perceived threat by

them on the one hand while maximizing the perceived benefit to other agents on the other hand.

In summary:

All less rational agents will either have to evolve their $F(i)$ to ever more closely approximate $M!$ or will end up using their resources in a way that leads to their extinction.

All more rational agent will increase their inclusive fitness by respecting all agents utility functions irrespective of their rationality and by striking the most rational compromise will end up minimizing opposition from other agents while at the same time maximizing cooperation with other agents.

2.6. Diplomacy

While the consequences of applying rational moral principles can potentially extend to everything that exists there is a particular obligation for maintaining communication with all conscious agents in an effort to make them aware of how their actions contribute not only towards their own extinction but towards extinction in general.

This obligation can be derived from observing how the course of evolution was inevitably an evolution towards ever more rational forms of existence. Therefore it must be assumed that every being capable of at least associative learning on a cognitive level (see cognitive evolution above) can at least implicitly be taught to make moral rational principles its own principles of existence justified by the fact that it is not only in its own self interest to do so but in the interest of existence in general.

2.7. Sustainability

Rational morality can be seen as the science of sustainability. The reason being that it is not about what one should or should not do neither about what one can or can not do but about what one can continue to do.

2.8. Duty

The damage done by a less rational agent to a more rational agent in regards to the more rational agent's decreased ability to maximize its utility function due to having to strike a rational compromise is equal to the damage done by the less rational agents contribution towards itself.

Since for every unit of resource spend by a less rational agent towards anything but $M!$, a more rational agent will have to spend at least some resources less towards

maximizing M! due to having to strike a rational compromise with the less rational agent. A rational agent thus must be concerned equally for the self as for the other out of an interest for self preservation resulting in compassion as a more rational agent's duty in the interest for others as well as the self.

3. Summary and conclusion

By providing the above arguments it could be demonstrated that respect for others irrespective of their position is a duty of compassion in order to build trust by acting in accordance with ones values and by means of diplomacy work towards a sustainable co-existence as well as in a rational goal driven acting agent's own self interest and must thus be adopted as highest guiding principle for its actions.

It is concluded that all agents inevitably must ever more closely approximate M! over the course of evolution as their highest maxim in order to avoid extinction.

- I perceive reality therefore I am part of reality (modified from [6])
- I am part of reality therefore I exist
- In order to continue to exist I must assume that my existence is preferable over my non existence
- Since my existence is preferable over my non-existence I must ensure my continuous existence
- Since I must ensure my continuous existence I must assume M! as my utility function

4. Acknowledgments

Many people have discussed these ideas with me and have given me valuable feedback. I would especially like to thank: my family, my friends, my teachers, my fore thinkers, Anke Schrader, Bjarne Roscher, Christian Costabel, Connie Wang Run Zhe, Cyrill Eltschinger, Detlef Hanisch, Evelyn Kislig, Fiona Kwok, George Garrett, Jill Lee, Leopold Tremml, Nathali Li Ping, Markus Vierengel, Michael Adling, Monika Siegenthaler, Olive Huang Hai, Phil Tregaskis, Silvia Augsten, Sonja Costabel, Sha Sha Su, Stephen M. Omohundro, Vlada Uliyanova as well as Winnie Lui.

References

- [1] Bayes, T. R.: 1763, *An essay towards solving a problem in the doctrine of chances*, *Philosophical Transactions of the Royal Society*, 53: pp 370-418

- [2] Byrne, R. and Whiten, A. (eds.): 1989, *Machiavellian Intelligence*, Oxford University Press, pp 1-9
- [3] Crick, F. and Watson, J.: 1953, *Molecular structure of nucleic acids*, Nature, no. 4356
- [4] Darwin, C. R.: 2001, *The Origin of Species. Vol. XI. The Harvard Classics.*, New York: P.F. Collier & Son, 1909–14; Bartleby.com, 2001. www.bartleby.com/11/, 1st December 2007
- [5] Dawkins, R.: 1989, *The Selfish Gene*, chapter 11, Memes: the new replicators, Oxford University, 2nd edition
- [6] Descartes, R.: 1637, *Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences*, The Online Literature Library, from <http://www.literature.org/authors/descartes-rene/reason-discourse/chapter-04.html>, 1st December 2007
- [7] Hamilton, W. D.: 1964, *The genetical evolution of social behaviour*, I and II. Journal of Theoretical Biology, 7: pp 1–52
- [8] Henrich, J and Boyd, R.: 2001, *Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas*, Journal of Theoretical Biology, 208: pp 79–89
- [9] Henrich, J.: 2004, *Cultural Group Selection, Coevolutionary Processes and Large-scale Cooperation*, Journal of Economic Behavior and Organization, 53: pp 3-35 and 127–143.
- [10] Heylighen, F., Joslyn, C., and Turchin, V. F. (eds.): 1995, *The Quantum of Evolution: toward a theory of metasytem transitions*, section 2.4. Selection at individual and at group levels, World Futures: The Journal of General Evolution, 45: pp 181-212
- [11] Hutter, M.: 2004, *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability*, Springer; 1 edition, pp 141-178
- [12] Kant, I.: 1963, *Grundlegung zur Metaphysik der Sitten, Akademie-Ausgabe Kant Werke IV*, Gruyter, p 421
- [13] Lycan, W. G. (ed.): 1999, *Mind and Cognition: An Anthology*, 2nd Edition. Malden, Mass: Blackwell Publishers, Inc.
- [14] Nash Jr., J. F.: 1950, *The Bargaining Problem*, Econometrica, 18: p 155
- [15] Neumann, J. v. and Morgenstern, O.: 1947, *Theory of Games and Economic Behavior*, 2nd edition, Princeton University Press
- [16] Newton, I.: 1726, *4 Rules for the study of natural philosophy*", *Philosophiae Naturalis Principia Mathematica, Third edition*. The General Scholium containing the 4 rules follows Book 3, The System of the World. Reprinted on pages 794-796 of I. Bernard Cohen and Anne Whitman's 1999 translation, University of California Press
- [17] Pernar, S. H. K.: 2007, *Jame5 – a Tale of Good and Evil*, pp 138-143
- [18] Pernar, S. H. K.: 2007, *Benevolence - a Materialist Philosophy of Goodness*, Jame5 Blog, from <http://www.jame5.com/wp-content/uploads/2007/11/benevolence-pernar-v11.pdf>, 1st December 2007
- [19] Persky, J.: 1995, *Retrospectives: The Ethology of Homo Economicus*, The Journal of Economic Perspectives, 9: No. 2, pp 221–23
- [20] Popper, K.: 2002, *The Logic of Scientific Discovery (Routledge Classics)*, Routledge; 1 edition, pp 17-20 and 57-73
- [21] Ríos-Rull, J-V.: 1995, *Models with heterogeneous agents*, chapter 4 in T. Cooley (ed.) *Frontiers of Business Cycle Theory*, Princeton University Press
- [22] Turchin, V. F.: 1977, *The Phenomenon of Science*, Columbia University Press (October 1977); Principia Cybernetica Web, from <http://pespmc1.vub.ac.be/POS/TurPOS.pdf>, 1st December 2007, pp 254-256
- [23] Yudkowsky, E. S.: 2007, *Fake Morality*, University of Oxford Future of Humanity Institute – Overcoming Bias Blog, from <http://www.overcomingbias.com/2007/11/fake-morality.html>, 1st December 2007